eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Tesis Doctoral

Reconstrucción del Paisaje de Waddington mediante algoritmos genéticos con análisis *Big Data* de datos transcriptómicos

Autor: Javier Cabau Laporta

Director: Marcos J. Araúzo Bravo

Programa de Doctorado en Biología Molecular y Biomedicina

Universidad del País Vasco / Euskal Herriko Unibertsitatea

2024

”Each individual can only hope by his efforts to come a little nearer than his predecessors to a full comprehension of the processes he is studying.”

Conrad Hal Waddington, Science Attitude (1948)

Índice general

Glosario de términos

ADN: *DNA*. Ácido desoxirribonucleico.

GRN: *Gene Regulatory Network*.

scRNA-seq: *single-cell RNA-Sequentiation.*

CELDA: *Cell: Expression, Localization, Development, Anatomy*.

CL: *Cell Line* (ontología).

CLO: *Cell Line Ontology.*

EFO: *The Experimental Factor Ontology.*

GEO: *Gene Expression Omnibus*.

OBO: *Open Biomedical Ontologies*.

ID: Es el identificador de cada clase en el contexto de las ontologías.

LMHA: *LungMap Human Anatomy*.

OWL: *Ontology Web Lenguage*.

TFs: Factores de transcripción.

TF: Verdaderos negativos.

TP: Verdaderos positivos.

FP: Falsos positivos.

FN: Falsos negativos.

OAEI:*Ontology Alignment Evaluation Initiative*.

HCA: *Human Cell Atlas*.

HTML: *HyperText Markup Language*.

GSE: *GEO DataSet Series*.

GSM: *Sample accessions numbers of GEO*.

GPL: *GEO platform.*

NCBI: *National Center for Biotechnology Information*.

RMA: *Robust Multichip Array*.

PCA: *Principal Components Analysis.*

vc: Cromosoma virtual.

td: Datos transcriptómicos.

vg: Gen virtual.

O: Ontología.

gr: Generación.

ct: Tipo celular.

rel: Relación de descendencia-ascendencia entre dos tipos celulares.

EC: Condición de salida del algoritmo genético.

DB: Decrecimiento básico.

DC: Decrecimiento creódico.

DEC: Decaimiento exponencial creódico.

GSEA: *Gene Set Enrichment Analysis*

GO: *Gene Ontology*.

Capítulo 1

Introducción general

El cuerpo humano, consiste en un conjunto de células que comparten el mismo ADN, pero que muestran diferentes morfologías y fenotipos a nivel individual. Las células conforman mediante diferentes asociaciones los órganos y tejidos que componen el cuerpo humano. La totalidad de las células que componen el organismo provienen de una única célula, el cigoto, la cual mediante una sucesión de divisiones y diferenciaciones irá generando los órganos y tejidos, compuestos por sus diferentes tipos celulares.

A lo largo de la diferenciación, las células pasarán de ser totipotentes (potencialmente capaces de generar individuos) a ser pluripotentes (células con capacidad de generar los distintos tipos de tejidos y órganos) y, a continuación, multipotentes (células más especializadas con capacidad de diferenciación en un determinado tejido o tipo celular). Finalmente, alcanzarán el grado final de diferenciación 'sacrificando' en el proceso la potencialidad celular para conseguir una especialización en una función concreta dentro del organismo.

Aunque entendemos y explicamos el proceso de diferenciación celular como una serie de secuencias y categorías, la realidad nos sugiere que las células, a lo largo del proceso de especialización son un *continuum*. Es decir, existen numerosos estadios intermediarios sin catalogar. Esto es fácilmente observable al comparar dos ontologías diferentes enfocadas en tipos celulares (Seltmann et al., 2013); (Edgar et al., 2013); (Cabau-Laporta et al., 2021), y al descubrimiento de nuevos tipos celulares fruto del uso de tecnología *single-cell* la cual genera datos masivos y supone el establecimiento y creación de Atlas de tipos celulares de diferentes organismos (Fei et al., 2022). El avance de estas tecnologías nos abre la posibilidad de descubrimiento de nuevos tipos celulares, los cuales nos permitiría descubrir nuevos tipos celulares 'inducibles' que no se dan en la naturaleza, pero pueden ser inducidos como intermedio entre dos tipos celulares. Para ello es ne-

cesario hacer hincapié en la generación de software orientado al *big data* que permita el análisis y obtención de información de la tecnología *single-cell* y la construcción de atlas como el *Human Cell Atlas* (Rozenblatt-Rosen et al., 2017).

En la actualidad, uno de los principales intereses a la hora de comprender los procesos de diferenciación celular es la posibilidad de inducir de forma ambiental un cambio en las células de modo que estas modifiquen su estadio a uno con mayor potencialidad o a otro con diferente especialización, a este proceso se lo conoce como reprogramación celular.

Una de las grandes contribuciones al campo de la reprogramación fue el descubrimiento del cocktail de factores de transcripción de Yamanaka (Takahashi, Yamanaka, 2006), que desarrolló un protocolo con el cual eran capaces de reprogramar células diferenciadas a estadios embrionarios mediante la introducción de determinados factores de transcripción (Oct3/4, Sox2, c-Myc y Klf4) en las células, induciendo su expresión (Takahashi, Yamanaka, 2006). El estudio de nuevas técnicas de reprogramación celular puede ayudar a mejorar las terapias aplicadas a diferentes afecciones, como puede ser el uso de autotransplantes del mismo paciente reprogramando, por ejemplo, células de su piel para sustituir células de otros tejidos que no estén sanas (Takahashi, 2012). Para el descubrimiento de nuevas formas de reprogramación celular, o para potenciar las ya existentes, es necesario comprender como funciona en detalle el proceso natural de diferenciación celular. Una forma de facilitar el avance en el conocimiento de los mecanismos subyacentes a la diferenciación celular pasa por la creación de un modelo computacional/matemático a partir de datos biológicos relacionándolos con la diferenciación celular.

En esta línea, se está revisitando la idea del paisaje epigenético de Waddington. El concepto inicial del paisaje epigenético de Waddington se extrae del trabajo realizado por Conrad Hal Waddington en la década de 1940, antes de ligar la herencia genética al ADN (Rajagopal, Stanger, 2016). Conrad Hal Waddington estudiaba la transmisión de caracteres adquiridos en la mosca de la fruta (Noble, 2015). Waddington realizó un experimento sobre cómo diferentes factores ambientales podían afectar al cigoto de la mosca de la fruta generando adultos con un tamaño diferente de alas y abdomen. A continuación, tras repetir el proceso un número de generaciones, las moscas mantenían la alteración aunque no recibieran el estímulo ambiental, al o que llamó asimilación genética (Noble, 2015).

Al conjunto de factores que generan la asimilación genética, Waddington lo llamó 'Epigenética' (*en o sobre* la genética) (Waddington, 1942) por lo que es conocido como el padre de la epigenética (Ingram, 2019). Waddington ideó en este contexto su diagrama, el paisaje epigenético (**Fig. 1.1**), como forma de explicar este proceso de desarrollo guiado por la epigenética. Para Waddington, su paisaje epigenético representaba la diferenciación del cigoto en los diferentes órganos, es decir, el paisaje de Waddington representaba únicamente el desarrollo desde cigoto hasta los órganos

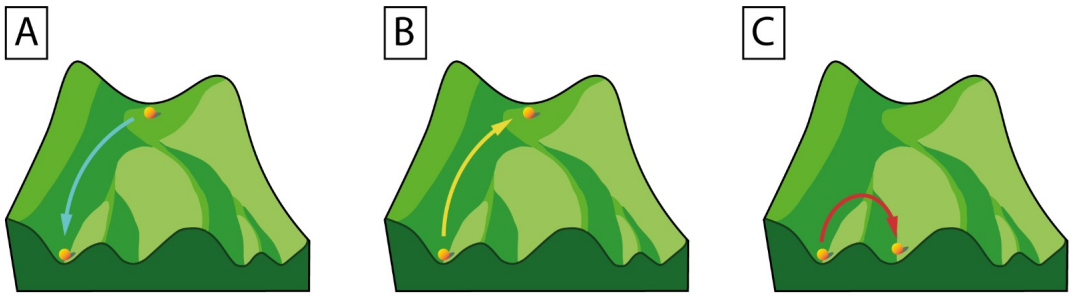


Figura 1.1: **Diagrama del paisaje epigenético de Waddington**. La bola representa en cada una de las figuras el estadio de diferenciación de una célula. A: Proceso de diferenciación celular, la bola cae del punto más alto al más bajo. B: Proceso de dediferenciación celular, la bola va del punto bajo hacia el más alto. C: Proceso de transdiferenciación celular, la bola va de un valle (mínimo local) a valle (otro mínimo local). Figura inspirada en 1 de Takahashi (2012)

completos (Fard et al., 2016). En él los puntos más elevados corresponden al cigoto y los valles finales a los diferentes órganos (Rajagopal, Stanger, 2016). Una bola que rodara por esos valles siempre realizaría los mismos recorridos, lo que él llamó 'creodos' (Rajagopal, Stanger, 2016), hasta llegar a su destino.

En el contexto de su investigación, Waddington postuló que el paisaje epigenético podía sufrir lo que denominó 'canalización genética' que es la alteración del paisaje epigenético por factores ambientales (epigenéticos), ya sean inducidos o alterados. La canalización la explicó como una 'persuasión' que se realiza sobre los cigotos para alterar el resultado del desarrollo del individuo (Trapnell, 2015), teniendo una tendencia a los flujos de desarrollo originales. Esta tendencia a volver a los flujos originales la denominó como homeorhesis para distinguirlo de la homeostasis de las células adultas (Rajagopal, Stanger, 2016). Para C.H. Waddington la influencia de la epigenética era una de las cuatro fuerzas que guían e influyen en la evolución de las especies (**Fig. 1.2**), en oposición a la teoría evolutiva común del momento basada en la genética de poblaciones (Loison, 2022).

Por un lado C. H. Waddington nos presenta un diagrama estático del desarrollo embrionario el cual puede modificar, generando así cierto dinamismo (Loison, 2022), la canalización genética antes mencionada. Inicialmente contemplaba este dinamismo como situaciones excepcionales, para al final presentar un nuevo añadido a su diagrama (**Fig. 1.3**), el cual dotaría al paisaje epigenético del dinamismo necesario para completar su teoría (Waddington, 1957). En este nuevo diagrama se presentarían una serie de cables que mediante su activación, por parte de la epigenética provocarían diferentes deformaciones en el paisaje epigenético y modificaría así el transcurso del desarrollo celular, generando al final distintos fenotipos en función de la epigenética (**Fig. 1.3**).

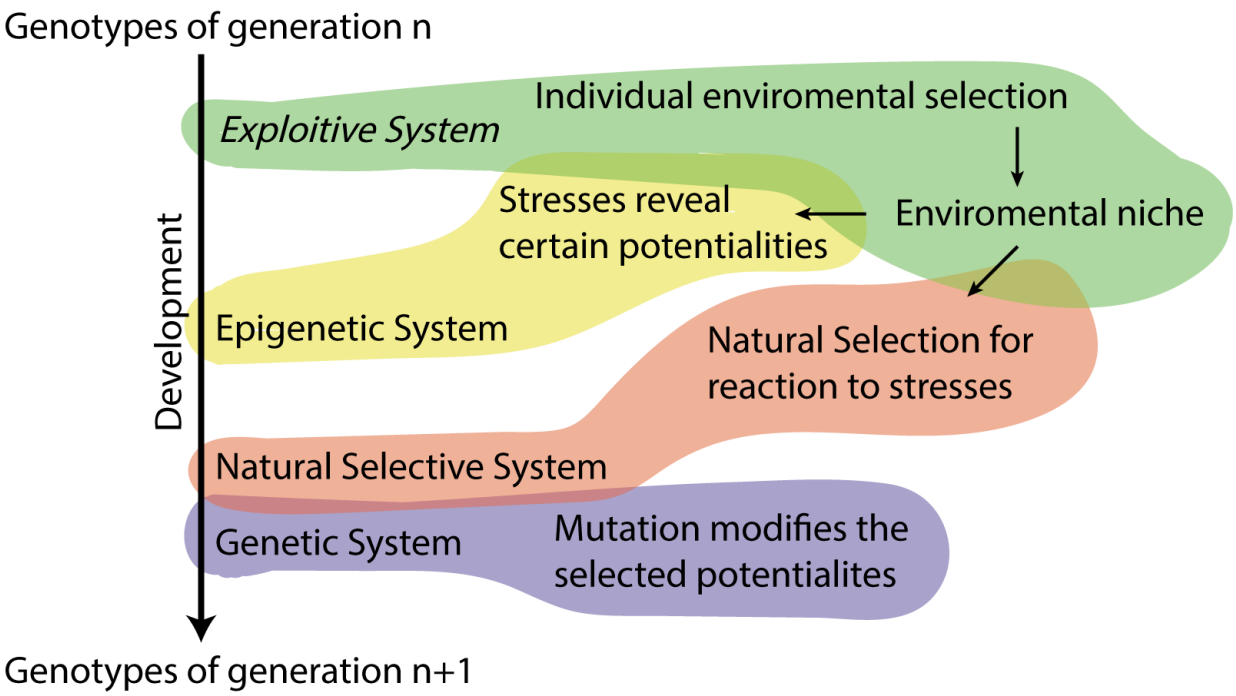


Figura 1.2: **Esquema de los sistemas de C. H. Waddington**. Cada uno de los sistemas descritos interactúa con el organismo influyendo a la adaptación del genoma de una especie (Waddington, 1959).

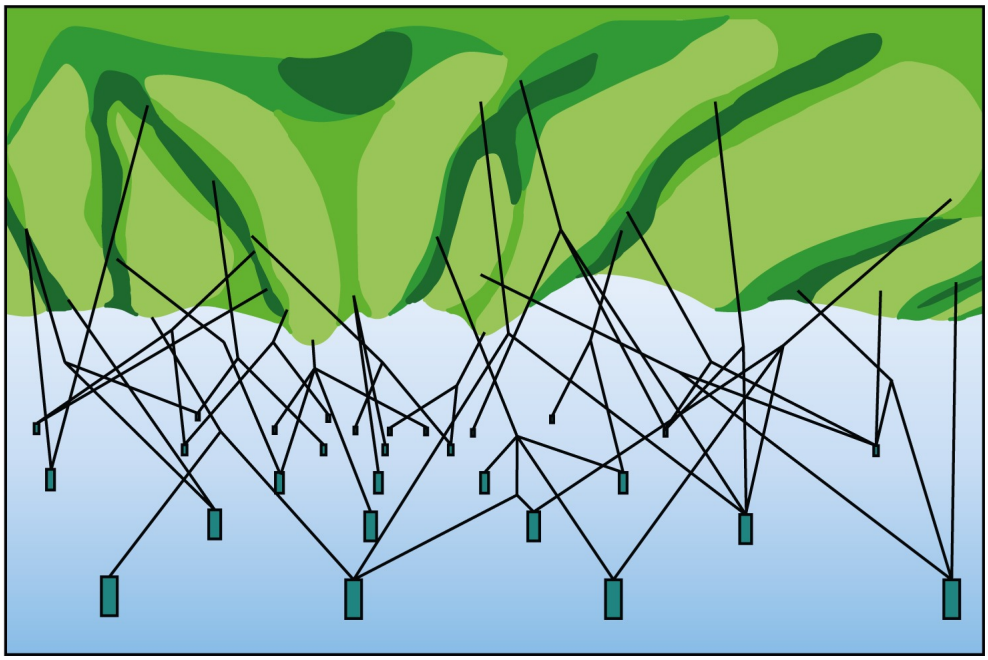


Figura 1.3: **Red Cibernética de Waddington**. En la red cibernética de Waddington se ilustra como unos cables podrían influenciar en la topografía del paisaje si se tensaran, generando los diferentes creodos en función de la activación de estos cables. Adaptada de (Waddington, 1957).

Recapitulando, C. H. Waddington ideó una teoría para explicar los cambios introducidos en las especies dotando de cierta influencia al entorno durante la embriogenesis (Loison, 2022). En este aspecto, un paisaje epigenético de Waddington podría interpretarse como la 'fotografía' del desarrollo embrionario de un individuo bajo unas condiciones *epigenéticas* concretas.

Más adelante, el paisaje epigenético de Waddington ha sido reinterpretado, y en la metáfora de la bola, esta ha sido asociada a una célula concreta eligiendo su camino en la diferenciación hasta un tipo celular totalmente diferenciado, en contraposición a la idea original de Waddington, en la que la bola representaría al cigoto y los valles finales los diferentes órganos (Noble, 2015);(Rajagopal, Stanger, 2016). En este caso, se comprendería el paisaje de Waddington, no como el crecimiento y desarrollo del cigoto si no como el crecimiento y variación de fenotipo de las diferentes células hasta conformar el organismo completo (Fard et al., 2016). Adicionalmente, rescata los conceptos establecidos por Waddington en su metáfora del paisaje epigenético, centrándose en la capacidad de diferenciación, respuestas correctivas y de reparación del organismo en lugar de la visión evolutiva original de Waddington (Rajagopal, Stanger, 2016). En este contexto de entendimiento del paisaje de Waddington se recogen los fenómenos de 'plasticidad celular' los cuales consisten en la capacidad natural de las células para pasar de un tipo celular a uno más indiferenciado o incluso a otro tipo celular (Rajagopal, Stanger, 2016).

En esta nueva visión dejamos a un lado la interpretación del paisaje epigenético de Waddington, aplicado a un conjunto de células en la construcción del organismo adulto y por tanto su respuesta evolutiva durante el desarrollo, para centrarnos en los procesos que subyacen a una célula para elegir un camino u otro a lo largo de la diferenciación celular. Las diferentes formas de este paisaje, sus valles y por tanto sus diferentes creodos vendrían dados por cambios en la expresión génica de la célula (Trapnell, 2015);(Fei et al., 2022);(Fard et al., 2016);(Fard, Ragan, 2019). A la luz de esta nueva interpretación del paisaje epigenético de Waddington, en la Red Cibernética de Desarrollo (**Fig. 1.3**) los diferentes cables que regulan los creodos supondrían el conjunto de genes y/o factores de transcripción que guían la diferenciación celular (Rajagopal, Stanger, 2016).

En ambas interpretaciones el paisaje de Waddington siempre se construyó sobre la existencia de cierto dinamismo, o bien contemplando 'fuerzas' que modifiquen los creodos (proceso de canalización) o bien a través de cables que controlan el relieve (Rajagopal, Stanger, 2016);(Noble, 2015), por lo que el paisaje de Waddington podría ser modificado en respuesta a estímulos externos (Epigenética). Estas deformaciones serían fruto de respuestas ambientales y darán lugar a los fenómenos que experimentan las células, como la dediferenciación o la transdiferenciación. Los cuales son un proceso por el cual una célula regresa a un estadio más indiferenciado (dediferenciación) o bien cambia su tipo celular a otro diferente (transdiferenciación) (**Fig. 1.1**) (Merrell, Stanger, 2016), conservando su estructura epigenética, diferenciándose así de los tipos celulares finales 'naturales' (Ladewig et al., 2013).

El paisaje epigenético de Waddington es una construcción ontológica de la diferenciación celular teniendo en cuenta diferentes factores que afecten a la misma. Esto convierte a la metáfora del paisaje epigenético en una forma de explicación de los mecanismos subyacentes de la totalidad de la diferenciación celular.

El paisaje epigenético de Waddington conecta con el concepto de 'tipo celular' expuesto anteriormente, es decir, entendiendo la diferenciación celular como el *continuum* que sigue 'la pelota que rueda por la colina hasta detenerse en un valle'. Por tanto, nos puede resultar una herramienta útil para diseñar posibles 'rutas alternativas' y conocer qué factores ambientales inducir a una célula para que esta pueda 'recorrer' un camino diferente al marcado por la naturaleza. Es decir, puede ayudarnos a descubrir nuevos mecanismos de dediferenciación y transdiferenciación atendiendo a los descubiertos actualmente (Rajagopal, Stanger, 2016).

Con todo lo anterior, se ha planteado el establecimiento de una relación *quasi-potencial* del paisaje epigenético de Waddington (Fard, Ragan, 2019), el cual relacionaría la elevación en el paisaje con la pertenencia a un determinado tipo celular (Huang et al., 2007). Con la abstracción de la altura en el paisaje de Waddington podríamos definir los diferentes tipos celulares de un organismo.

En la actualidad, se entiende el paso de un estadio celular al siguiente como dependiente de determinados factores de transcripción (TFs) y de otros genes de la Red Reguladora Genética, *Genetic Regulatory Network*, (GRN) (Fard et al., 2016). Estos factores de transcripción interactúan entre sí, creando una red de activación-inhibición que causa una cascada de diferenciación, donde la célula puede alcanzar un número discreto de estados mutuamente excluyentes (Müller-Molina et al., 2012). En la metáfora del paisaje epigenético de Waddington presentando en *The strategy of the genes*;(Waddington, 1957) los TFs podrían entenderse como los diferentes cables que regulan la red Cibernética de Waddington (**Fig. 1.3**). La activación de las diferentes rutas de la GRN ocasiona la aparición de regiones del paisaje epigenético de Waddington que son 'habitualmente visitadas' durante la diferenciación celular, es decir, regiones de alta estabilidad donde tenderían a acumularse los diferentes fenotipos que atraviesa una célula durante su diferenciación, estas regiones son conocidas como atractores(Fard et al., 2016), y por definición, los tipos celulares plenamente diferenciados reciben este nombre en el contexto del paisaje de Waddington. Estos atractores pueden definirse como estados de equilibrio correspondientes a los fenotipos más probables (Fard, Ragan, 2019) por tanto la construcción de un paisaje de Waddington podría posibilitar el descubrimiento de nuevos atractores.

El proceso por el cual la GRN programa a la célula para seguir un camino u otro es conocido como multiestabilidad (*multistability*) (Wu et al., 2017). En el caso de organismos multicelulares, las células pueden utilizar la multiestabildiad para diferenciarse en otros tipos celulares, mientras que, en el caso de organismos unicelulares, como las bacterias, pueden utilizar la multiestabildiad

para cambiar entre diferentes estadios de respuesta al entorno (Wu et al., 2017). Esto también abriría la posibilidad para el estudio de paisajes de Waddington en el mundo procariota, relacionando los tipos celulares a los diferentes fenotipos que atraviesan las bacterias en respuesta al entorno (Sánchez-Romero, Casadesús, 2021), creando así un paralelismo con la dediferenciación y la transdiferenciación en la que células 'individuales' de un organismo pluricelular transforman su fenotipo en función de un cambio en el ambiente.

A día de hoy el mecanismo necesario para que una célula elija un creodo u otro sigue siendo desconocido (Wu et al., 2017);(Fei et al., 2022), aunque es posible que los Factores de Transcripción jueguen un rol importante en este proceso (Huang, 2009);(Ladewig et al., 2013), además de que también es posible que existan otras formulaciones en otros mecanismos para la toma de decisiones de la célula, en algunos casos intrínsecos: que incluiría los mecanismos de metilación de DNA, la modificación de histonas y las vías de señalización (Fard et al., 2016) y en otro extrínsecos que incluiría fuerzas externas como la gravedad o el magnetismo (Bizzarri et al., 2020), de hecho una de las mayores dificultades a la hora de simular organismos vivos es la de tener en cuenta todas estas fuerzas, tanto las extrínsecas como las intrínsecas (Bizzarri et al., 2020).

En la línea el estudio de los fenómenos de transdiferenciación y dediferenciación mediados por Factores de Transcripción, estos pueden darnos una pista sobre el funcionamiento de las 'decisiones' que toma la célula para elegir un destino u otro (Fard et al., 2016); (Takahashi, 2012);(Fei et al., 2022). Se han realizado varios avances en la comprensión y estudio de estos procesos como por ejemplo en respuesta a la expresión de distintos factores de transcripción (Takahashi, 2012). En el campo de la dediferenciación inducida nos encontramos con la pluripotencia inducida mediante los factores de transcripción Oct3/4, c-Myc, Sox2 y Klf4 a fibroblastos (Takahashi, Yamanaka, 2006), en el campo de la transdiferenciación inducida cabe destacar la reprogramación de células de la glia a neuronas (Heins et al., 2002), células del páncreas a células hepáticas (Shen et al., 2003), células B a macrófagos y células T (Xie et al., 2004);(Maherali et al., 2007), fibroblasto a macrófago (Xie et al., 2004), fibroblasto a cardiomicitos, células hepáticas, células hematopoiéticas, células condrogénicas, células dopaminérgicas y neuronas motoras espinales (Szabo et al., 2010);(Sekiya, Suzuki, 2011); (Son et al., 2011).

En la naturaleza, los mecanismos que operan en los fenómenos de dediferenciación y transdiferenciación son aún una incógnita (Riva et al., 2022) ambos fenómenos pueden observarse en la respuesta del organismo a la reparación de heridas y como un 'mal funcionamiento' en el caso de las células tumorales que adquieren cierto grado de pluripotencia (Merrell, Stanger, 2016).

Se han propuesto otros modelos metafóricos inspirados en el paisaje epigenético de Waddington que explican la diferenciación celular y la elección de 'destino' tal y como hace el paisaje de Waddington haciendo más énfasis en la plasticidad celular y los fenómenos de dediferenciación

y transdiferenciación que en los procesos de desarrollo y diferenciación celular natural. Este es el caso de 'Disco Epigenético', *Epigenetic Disc* propuesto en (Ladewig et al., 2013) (**Fig. 1.4**). Este modelo propone que las células pluripotentes actuarían como un *primus inter pares*, es decir, estarían al mismo nivel sin establecer jerarquías. Sería pues la acción externa la que realizaría modificaciones y decantaría el destino celular en un sentido u otro. En la metáfora del Disco Epigenético, la célula sería nuevamente una bola sobre un disco plano, las células diferenciadas se situarían en los extremos, aplicando fuerzas contrarias (los diferentes cambios ambientales) en distintos puntos del disco, y contando con ciertos raíles (TFs), se podría modificar la pendiente y guiar el recorrido de la bola hacia cualquiera de las células (Ladewig et al., 2013).

El modelo del Disco Epigenético resulta de mucha utilidad si focalizamos los procesos de cambio de fenotipo celular en los factores externos a la célula aplicados de forma puntual. Sin embargo, si pretendemos observar los cambios de forma secuencial en un organismo resulta más útil el modelo clásico del paisaje epigenético de Waddington, ya que en un paisaje Epigenético podemos representar todas las relaciones y en un Disco Epigenético (**Fig. 1.4**) solo podemos modelar una a la vez.

La importancia de la modelización/construcción de un paisaje de Waddington consiste en el estudio y profundización de los mecanismos que regulan la diferenciación celular en los distintos pasos y tendría aplicaciones directas, no solo en profundizar en el entendimiento biológico del desarrollo celular, sino en los fenómenos de diferenciación y transdiferenciación, ya que existe la hipótesis de que un estudio en profundidad sobre un paisaje de Waddington podría revelar la existencia de 'túneles' y 'subidas' (o modificaciones si interpretamos el paisaje de Waddington desde el punto de vista dinámico) utilizadas de forma natural para los mecanismos de dediferenciación y transdiferenciación (además del descubrimiento de 'túneles y subidas' alternativas) y que podrían extrapolarse en el laboratorio para desarrollar estrategias que puedan aplicarse en el laboratorio para reprogramación celular y aplicarlo más adelante en terapias (Takahashi, 2012); (Rajagopal, Stanger, 2016). También, resultaría útil estudiar las alteraciones al paisaje de Waddington que generaran afecciones como el cáncer o heridas en diferentes tejidos celulares que activen respuestas regenerativas de dediferenciación y transdiferenciación en células espacialmente próximas (Rajagopal, Stanger, 2016).

En un diagrama basado en la metáfora del paisaje epigenético de Waddington el eje z (o energía quasi-potencial (Huang, 2009)) sería el valor principal de la jerarquía, el cual nos ordenaría el grado de diferenciación y nos mostraría los diferentes creodos y atractores a lo largo del paisaje de Waddington. Abordando la determinación del eje z para la construcción del paisaje de Waddington con datos transcriptómicos y focalizándonos en predecir qué camino elegirá una determinada célula, se han utilizado tres estrategias diferentes, mediante métodos lógicos, métodos continuos y métodos no supervisados (Fard et al., 2016). Todos los diagramas tiene dos presuposiciones: en

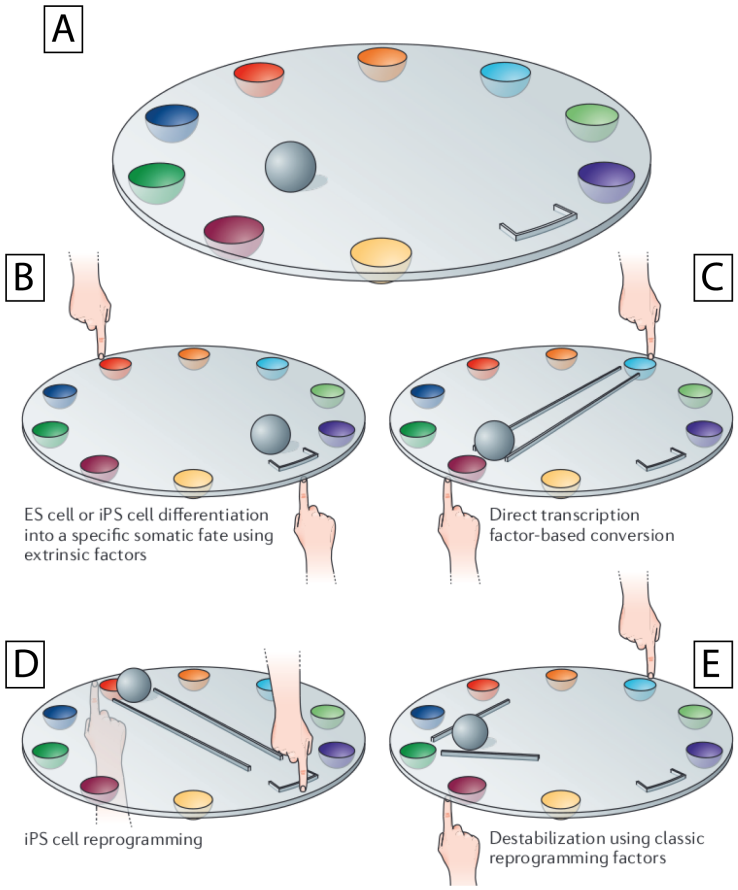


Figura 1.4: **Diagrama del Disco Epigenético**. En este diagrama podemos observar el funcionamiento de la metáfora. En esta metáfora al aplicar fuerzas opuestas (Epigenética/Factores extrínsecos a la célula), podríamos crear pendientes y dirigir la bola (tipo celular inicial) a uno de los diferentes pocillos (tipo celular final), también ayudándonos de railes para dirigir el movimiento de la bola (Factores de Transcripción). Se observan diferentes ejemplos de este modelo los cuales son: A visión general del modelo. B diferenciación de células ES o iPS a una célula somática específica. C conversión de tipos celulares mediante factores de transcripción. D reprogramación de células iPS. E desestabilización utilizando factores clásicos de reprogramación. Adaptación de: (Ladewig et al., 2013).

primer lugar, se considera que perfiles de expresión genética convergen hacia atractores similares, y en segundo lugar, que en una configuración determinada del GRN corresponde a un tipo celular concreto (Fard et al., 2016).

Por un lado, los métodos lógicos consisten en crear una red lógica basada en estados booleanos, es decir, estos métodos tienen en cuenta qué genes se activan y se desactivan en cada paso. Este método simplifica demasiado el esquema de un organismo y es posible que sea demasiado reduccionista. Por otro lado, los métodos continuos capturan mejor el dinamismo del sistema y se expresarían mediante ecuaciones diferenciales ordinarias (EDOs), en las que el resultado de la expresión de un gen tendría efectos de activación inhibición en otros. Al igual que en los métodos lógicos, los métodos continuos requieren información específica del sistema para su construcción (Fard, Ragan, 2019), los métodos lógicos requieren información de activación-desactivación de genes en cada paso, mientras que los métodos continuos requieren información de las constantes cinéticas de las reacciones, por lo que presentan un problema a la hora de escalar estos sistemas a sistemas complejos, como el de un organismo (Fard, Ragan, 2019). Existen diversas aproximaciones a la construcción del paisaje de Waddington con pocos elementos como es el caso de Huang (2009), donde utilizaron un método continuo para simular las dinámicas de un paisaje de Waddington de tres tipos celulares y regulada por la activación-inhibición de dos factores de transcripción (GATA1 y PU.1) (Huang, 2009).

En contraste con estos dos métodos, tenemos los métodos no supervisados, los cuales mediante el uso de algoritmos son capaces de modelar un paisaje de Waddington sin conocimiento previo de las dinámicas del sistema. Por tanto, estos métodos no están sujetos a las limitaciones de los anteriores (Fard et al., 2016).

En la línea de los métodos no supervisados para la construcción del paisaje de Waddington, hay que destacar el uso de las Redes Hopfield por parte de Fard et al. (2016) y Fard, Ragan (2019). Las Redes Hopfield son un tipo de red neuronal utilizada principalmente para el reconocimiento de patrones, las cuales utilizan una función de energía. Esta se utiliza de forma análoga a la energía quasi-potencial de Waddington en Fard et al. (2016) y Fard, Ragan (2019). Esta función de energía se minimiza cuando la red alcanza un estado estable, los cuales se identificarán como atractores. De esta forma es capaz de identificar los múltiples atractores a partir de datos transcriptómicos de un conjunto de tipos celulares, generando jerarquías y capturando el desarrollo celular (Fard et al., 2016).

Si bien, las Redes Hopfield pueden construir un paisaje de Waddington estático (Fard, Ragan, 2019). En nuestro caso optamos por el uso de algoritmos genéticos, ya que la propia lógica del Algoritmo Genético permite mayor facilidad para implementar un sistema en paralelo que ayude con la excesiva carga computacional que ambos sistemas generan (Whitley, 1994).

Por un lado el uso de Redes Hopfield aplicadas a este problema requiere un *training set* y una selección previa de genes, ya que la exactitud (*accuracy*) podría verse comprometida si el numero de *features* introducido es muy amplio (fenómeno denominado como: 'La maldición de la dimensionalidad') (Fard, Ragan, 2019), para el Algoritmo Genético necesitamos la entrada de una jerarquía de desarrollo celular (una ontología que establezca las bases y relación del paisaje de Waddinton), la elaboración de una función objetivo (aquella función que nos genere un mejor ajuste de la expresión génica con la diferenciación) y un ajuste previo de parámetros (que garanticen la mejor convergencia del algoritmo genético), aunque también es aconsejable una inicialización, es decir, una reducción previa de los *features* introducidos para reducir la carga computacional.

En el presente trabajo, vamos a desarrollar las herramientas, y análisis necesarios para la construcción de un paisaje de Waddington utilizando datos transcriptómicos humanos y su posterior evaluación. Utilizaremos como energía *z quasi-potencial* los datos transcriptómicos alojados en repositorios online y utilizaremos un algoritmo genético como medio para seleccionar los genes que expliquen mejor el paisaje de Waddington para el conjunto de datos que obtengamos.

En primer lugar, necesitaremos una relación ontológica basada en criterios biológicos de diferenciación de los distintos tipos celulares que conformen el organismo humano. CELDA (Seltmann et al., 2013) y LifeMap (Edgar et al., 2013) son en la actualidad dos de las ontologías más completas que recojan información sobre la diferenciación celular. Por ello, el primer paso será elaborar una ontología consenso a partir de ambas.

A continuación, una vez conozcamos las diferentes clases de esta nueva ontología consenso, procederemos a descargar los datos disponibles para el conjunto de tipos celulares, utilizando algoritmos automáticos para la descarga masiva de datos. En este punto será necesario desarrollar el software para el etiquetado y descarga de los datos alojados en GEO basándonos en los metadatos.

Para la aplicación del algoritmo genético es necesario el establecimiento de la 'función objetivo' la cual debe reflejar y otorgar puntuación a los datos introducidos (Whitley, 1994). Esta función objetivo, será ideada para seleccionar los genes que construyan el paisaje de Waddington. El siguiente paso para la aplicación de algoritmos genéticos es la evaluación de convergencia (un no-estancamiento en mínimos locales) del algoritmo. Para ello, elaboraremos una prueba de Concepto con un grupo reducido y manejable de tipos celulares, esto nos permitirá elaborar el software y realizar el ajuste de parámetros, también nos permitirá evaluar el grado de funcionamiento de los algoritmos existentes de reducción del *Batch Effect*.

El último paso del algoritmo genético consistirá en la observación de respuesta de escalado desde la prueba de concepto a la totalidad de los datos obtenidos con las herramientas de descarga y

etiquetado masivo. Se procederá a la selección de los genes que permitan mejor la construcción del paisaje de Waddington y finalmente, se estudiará la naturaleza de los genes obtenidos en el algoritmo genético.

Las hipótesis de la presente tesis son:

- **Hipótesis 1**: La utilización de datos transcriptómicos podría revelar la estructura subyacente del paisaje de Waddington, proporcionando una visión detallada de los estados celulares y sus transiciones durante el desarrollo.
- **Hipótesis 2**: La recopilación de datos transcriptómicos de fuentes públicas podría permitir una construcción del paisaje de Waddington.
- **Hipótesis 3**: La identificación y evaluación de genes relacionados con la construcción del paisaje genético a través de algoritmos genéticos podría revelar los genes clave involucrados en la determinación de trayectorias celulares, y también los mecanismos biológicos subyacentes a estos procesos de diferenciación y desarrollo.

Para probar estas hipótesis se plantean los siguientes objetivos:

- **Objetivo 1**: crear un software para el alineamiento y *merging* (fusión) de ontologías, el cual nos permitirá el establecimiento de ontologías consenso, aplicado en el presente trabajo para generar una ontología con el máximo posible de información sobre desarrollo celular.
- **Objetivo 2**: elaborar algoritmos de *scrap web* para el etiquetado y descarga de datos masivos de GEO.
- **Objetivo 3**: crear el software asociado al algoritmo genético, de modo que este pueda ser implementado en el pipeline y se adapte a las necesidades de este trabajo (tipo de dato, función objetivo, etc). También, establecer de una 'función objetivo' que permita seleccionar aquellos 'rasgos' que construyan un paisaje de Waddington y sea abordable a nivel computacional por un algoritmo genético.
- **Objetivo 4**: construir un paisaje de Waddington con datos transcriptómicos, reciclados de datos públicos en una base de datos.
- **Objetivo 5**: observar y evaluar los genes obtenidos mediante el algoritmo genético para la construcción del paisaje genético y buscar los mecanismos biológicos relacionados.

Capítulo 2

Estandarización de bases de datos para la construcción de una ontología de desarrollo celular

2.1. Introducción

El avance de las tecnologías biomédicas de precisión origina que se produzcan cada vez mayor cantidad de información ómica, no sólo a nivel de datos aglomerados, sino también de a nivel unicelular (Hwang et al., 2018) e incluso subcelular (Grindberg et al., 2013) que permiten descubrir nuevos tipos celulares (Boldog et al., 2018); (Gerovska, Araúzo-Bravo, 2019); (Sas et al., 2020).

Estos datos celulares, cada vez más detallados, ha provocado que los antiguos sistemas de clasificación celular queden obsoletos y crean una demanda de nuevos métodos automáticos de clasificación celular. Una de las estructuras para clasificar elementos de un dominio del conocimiento, como por ejemplo los tipos celulares, son las ontologías. Estas se pueden definir de varias maneras según el contexto de uso (Busse et al., 2015), en ciencias de la información, una ontología se define como una tupla de siete elementos, $O := L, C, R, F, G, T, A$, donde $L := LCLR$ es un léxico de conceptos LC y relaciones LR; C es un conjunto de conceptos; R es un conjuntos de relaciones binarias en C; F es una función que conecta conceptos de símbolos a conjuntos de conceptos $Sub(LC) \rightarrow Sub(C)$; G es una función que conecta relaciones de símbolos a conjuntos de relaciones, $Sub(LR) \rightarrow Sub(R)$; T es una taxonomía para el ordenamiento parcial de C, T(Ci, Cj), y A es un conjunto de axiomas con elementos C y R (Busse et al., 2015). Una pregunta crítica a responder durante el diseño de la ontología es el nivel de detalle que debe explicar dicha ontología, dicho nivel de detalle se denomina granularidad. Así, diferentes ontologías del mismo

dominio del conocimiento utilizan diferentes conceptualizaciones para obtener el nivel de granularidad deseado. En el caso de las ontologías celulares, existen varias clasificaciones de tipos celulares en varios formatos siendo el más utilizado el *Ontology Web Lenguage* (OWL) (Antoniou, Harmelen, 2009), que abarca la gran mayoría de las ontologías de OBO Foundry (*Open Biomedical Ontologies*) (Smith et al., 2007).

El creciente número de nuevos tipos celulares descubiertos impulsados por generación de datos como *single-cell* RNA-Seq (scRNA-seq) e iniciativas internacionales como Human Cell Atlas (HCA) (Rozenblatt-Rosen et al., 2017) crea una necesidad desarrollar métodos computacionales para facilitar la automatización de la creación de ontologías celulares y la clasificación de estas nuevas células como ramas de ontologías celulares existentes (Osumi-Sutherland, 2017).

Se pueden crear nuevas ontologías celulares reutilizando y fusionando la información dispersa en múltiples ontologías celulares. Antes de fusionar (*merge*) dos ontologías, es necesario encontrar las correspondencias entre sus conceptos en un proceso llamado alineación de ontología (*ontology alignment*) o emparejamiento (*ontology matching*). Existen numerosas herramientas para la alineación y fusión de ontologías (**Tabla 2.1**). La mayoría de estas herramientas son semiautomáticas ya que requieren una entrada inicial y algunas entradas intermedias del usuario para realizar la alineación. Algunas de estas herramientas solo se enfocan en la alineación de ontologías y aunque es un paso esencial en el proceso es importante la incorporación de la fusión a partir de esa alineación.

Para minimizar la supervisión humana de la alineación de ontologías y automatizar la fusión de ontologías, desarrollamos un algoritmo y lo implementamos en FOntCell Cabau-Laporta et al. (2021), un paquete de software en Python para la fusión automática de ontologías. Aplicamos FOntCell para crear una nueva ontología más completa y detallada del desarrollo celular mediante la fusión de ontologías celulares de todos los tipos celulares del cuerpo humano.

Existen varias ontologías que contienen diferente información biomédica (genómica, proteómica y anatómica) (Lambrix et al., 2007). Las dos ontologías más grandes que aportan información en cuanto a diferenciación y desarrollo celular son CELDA (Seltmann et al., 2013) y LifeMap (Edgar et al., 2013). CELDA integra información biomédica diversa, tal como: expresión genética, localización celular, desarrollo y anatomía, de datos *in vivo* como *in vitro*, tanto de células humanas como de ratón. La parte de CELDA relevante para la construcción del paisaje de Waddington es la anotación de desarrollo celular, información contenida en los campos de Cell Ontology (CL) (Bard et al., 2005), Cell line Ontology (CLO) (Sarntivijai et al., 2014) y Experimental Factor Ontology (EFO) (Malone et al., 2010).

Tool	Name Mapping	Structure Mapping	Auxiliary	Automatic	Merging
ArtGen	Name	Parents, children	WordNet	Semi or fully	
ASCO	Name, label, description	Parents, children, sibilings, path from root	WordnNet	Fully	
Chimaera	Name	Parents, children		Semi	Merging
FCA-Merge	Name			Semi	Merging
FOAM	Name, label	Parents, children, Equivalence		Semi	
GLUE	Name			Semi	
HCONE	Name	Neighborhood	WordNet	Semi	Merging
IF-Map		Parents, children	Reference ontology	Semi	
iMapper		Domain, range	WordNet	Semi	
Onto Mapper	Name	Parents, children		Semi	
Anchor-PROMPT	Name	Direct graphs		Semi	Merging
SAMBO	Name, synonym	is-a part-of, descendants & ancestors	WordNet UMLS	Semi	Merging
S-Match	Label			Fully	
AML	Label, instances	Direct graph, logical repair algorithm	WordNet	Fully	
LogMap	Label, name	Linguistic alignment, principle of locally	WordNet, UML-lexicon	Semi or fully	
AGM	Name, label	Graphs		Seni or fully	
ALIN	Label		WordNet	Semi	
DOME	Label	doc2vec		Fully	
FCAMap-KG	Label, synonym	Part-of		Semi or fully	
Lily	Name, label	Direct graphs		Semi or fully	
LogMapBio	Label, name	Linguistic alignment, principle of locally	WordNet, UMLS-lexicon, BioPortal	Semi or fully	
LogMapLite	Label, name		WordNet, UMLS-lexicon	Semi or fully	
POMAP++	Name, label	Ontology atribute, linguistic match		Semi	
FOntCell	Label, synonym	Direct graphs, atribute relation, ontology atribute, linguistic match		Fully	Merging

Tabla. 2.1: **Comparativa de las diferentes herramientas para el alineamiento y fusión de ontologías.** En la primera columna se muestra el nombre de la herramienta, en la segunda los argumentos que contempla para el alineamiento por nombre. En la tercera columna se muestran el tipo de alineamiento estructural que utiliza. La cuarta columna nos muestra el uso de recursos externos auxiliares para mejorar el alineamiento. La quinta columna nos indica si la herramienta es automática o semiautomática. En la sexta columna se muestra si la herramienta realiza el *merging* o fusion de las ontologias. Las celdas en blanco indican que la herramienta en cuestión no utiliza ese tipo de proceso

Por otro lado, existe LifeMap, que es un repositorio con importante información celular (Edgar et al., 2013). En LifeMap se incluye información de diferenciación y desarrollo celular. Ambas ontologías se complementan en cuanto a diferentes tipos celulares que están presentes en una de las dos y no en la otra.

Por lo general, diferentes ontologías utilizan diferentes etiquetas para un mismo tipo celular y un simple *word matching* no es suficiente para encontrar las equivalencias (Lambrix, Tan, 2008). Por lo que para una correcta alineación por nombre es necesario desarrollar nuevas estrategias que relacionen clases de ambas ontologías que no solo utilicen el etiquetado de las mismas si no estructuras internas de las ontologías para encontrar equivalencias.

Existen, también, otras ontologías de diferenciación y desarrollo de tipos celulares de tejidos más específicos como es el caso de LungMap Human Anatomy (LMHA) Ardini-Poleske et al. (2017) la cual es una ontología específica de células del tejido pulmonar. Estas ontologías podrían ampliar también el conocimiento específico de una ontología final, yendo a un nivel más detallado.

2.2. Materiales y métodos

Los principales pasos implementados en FOntCell a la hora de hacer la unión entre dos ontologías son:

- Introducción de documentos
- Análisis (gramatical) de las ontologías
- *Pre-processing* de las ontologías
- Alineamiento
- Fusión

Los parámetros que utiliza FOntCell para realizar el alineamiento se adjunta en un documento de configuración (**Fig. 2.1**) donde se especifican los mismos. Más adelante, FOntCell realiza el alineamiento buscando equivalentes, utilizando una combinación de emparejamiento por similitud de nombre (*name matching*) y por emparejamiento de similitud estructural/topología de grafos (**Fig. 2.2**). La fusión, finalmente, actúa manteniendo las relaciones comunes y añadiendo las relaciones no comunes al resultado final a partir de las equivalencias. En esencia, FOntCell busca las clases similares (para emparejarlo) y las clases diferentes (para añadirlo).

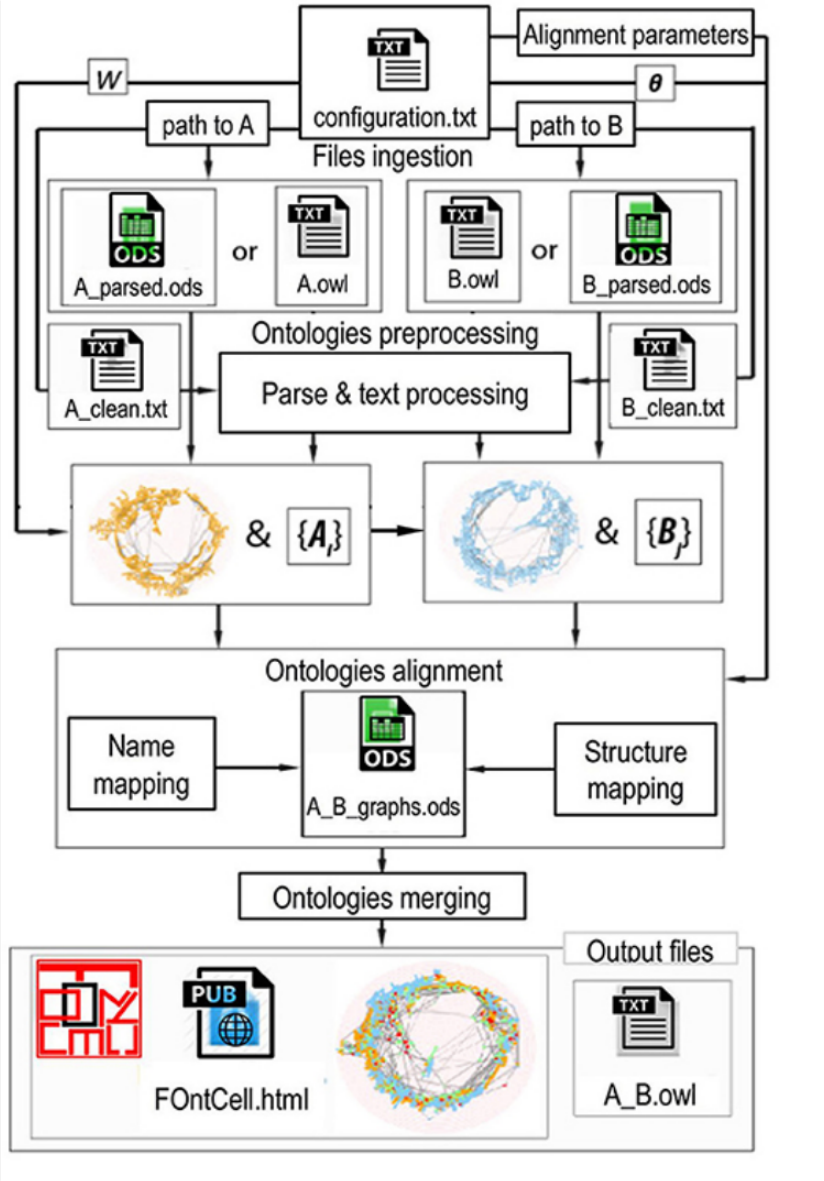


Figura 2.1: **Diagrama de flujo de FOntCell**. En este diagrama se muestran las principales funcionalidades, ingestión de documentos, preprocesado, análisis, alineamiento, fusión y generación de documentos output. Junto con las ontologías el usuario proporciona los valores como el ancho de ventana 'W' y los límites de similitud del vector $\Theta = \{\Theta_N, \Theta_T, \Theta_{LN}\}$

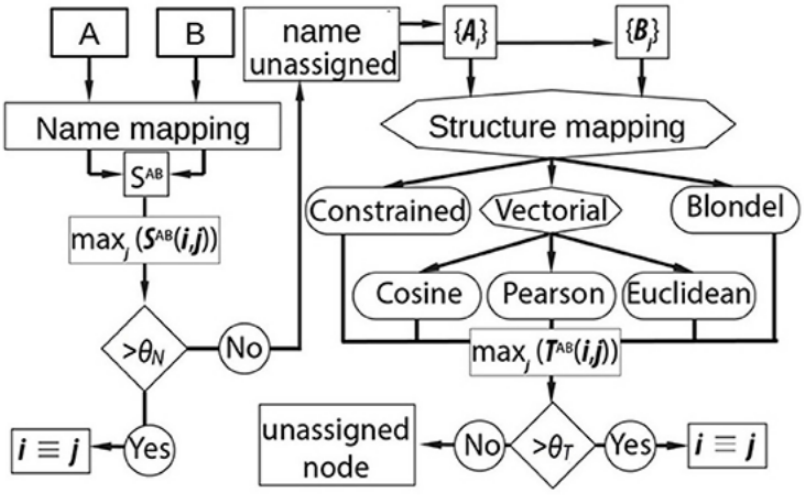


Figura 2.2: **Diagrama de flujo del algoritmo de alineamiento de FOntCell** Este diagrama combina el mapeo de nombres (*name mapping*) a la izquierda y el mapeo estructural (*structure mapping*) a la derecha. El mapeo estructural utiliza cinco métodos alternativos: *Constrained*, *Vectorial-cosine*, *Vectorial-Pearson*, *Vectorial-Euclidean* y *Blondel*. {A} y {B} hacen referencia al set de subgrafos alrededor de los nodos i y j de las ontologías A y B respectivamente.

2.2.1. Análisis de ontología

El Análisis de ontología (*ontology parse*) es esencial para computar la ontología. El 'análisis' hace referencia a su 'análisis gramatical', esto nos permite la lectura e interpretación de la ontología por parte de FOntCell, y por tanto poder trabajar con ella.

Esta lectura de las ontologías genera dos matrices de dos columnas, que denominamos A2 y B2, cada una con un número de filas igual al número de relaciones entre clases que tenga cada ontología. La primera columna tiene el nombre de cada clase y la segunda, el nombre de cada uno de sus clases hijas. Las matrices A2 y B2 son ficheros intermediarios necesarios para el mapeo estructural.

Hemos diseñado FOntCell para que tenga gran flexibilidad la fusión ontologías con distintos formatos tales como .owl, matrices A2 y B2 en formato tabulado .txt o bien documentos .ods (compatible con el módulo de python pyexcel-ods) que tengan un formato de matrices A2 y B2.

2.2.2. Pre-procesado de ontologías

Aunque dos ontologías pertenezcan al mismo ámbito del conocimiento no implica que la distribución en cuanto a clases y estructura sea la misma.

El pre-procesado es un paso opcional dentro del algoritmo de FOntCell. Este pre-procesado funciona permitiendo una 'reparación' de las ontologías entrantes, mediante la elección del tipo de relación del que extraer información (instancia, descendientes, ascendentes...), editar etiquetas y modificar las relaciones de la ontología principal por adición, borrado y fusión de clases y relaciones dentro de una misma ontología. Gracias al pre-procesado de ontologías con FOntCell podemos extraer la información de interés que luego nos servirá para hacer un correcto alineamiento y por tanto una fusión de ontologías. El pre-procesado también se encarga de eliminar diferentes fallas en las ontologías, eliminando clases duplicadas o reconectando ramas desconectadas.

En FOntCell el pre-procesado es automático, para dirigirlo se utiliza como entrada un documento .txt donde están escritas en forma de comandos las ordenes necesarias para su pre-procesamiento:

- Elegir clases por palabras en su etiqueta: por ejemplo, aquellas que contengan la palabra *cell* para descartar aquellas clases que se refieren a tejidos o a conceptos más abstractos.
- Eliminar clases: elimina clases que contenga una determinada palabra o bien introduciendo la ID de la clase. Por ejemplo, utilizado para eliminar aquellas clases de CELDA que incluyan la palabra '*cell line*', '*inmortalized*' o '*derived*'.

- Introducir nuevas clases: de ser necesario, es posible introducir una nueva clase añadiéndola en una relación con otra.
- Concatenar clases: Crea una nueva relación entre clases de la ontología.
- Fusionar clases: fusiona dos clases de una misma ontología, manteniendo las etiquetas como sinónimos y fusionando las relaciones de ambas clases. El resultado conservará la ID de una de las dos clases.
- Pre acondicionamiento de etiquetas: Editando los nombres, etiquetas y sinónimos de las clases. Se eliminan términos generales como: '*the*', '*cell*', '*cells*', '*human*', '*mouse*', o símbolos como guiones. Este paso permite un emparejamiento por nombre más específico.

	Antes del pre-procesado		**Después del pre-procesado**	
	#Clases	#Relaciones	#Clases	#Relaciones
CELDA	15.439	203.058	841	966
LifeMap	796	924	796	924
CELDA+LifeMap	1.408	1.855	-	-
LMHA	80	13	45	45
(CELDA + LifeMap) + LMHA	1.437	1.919	-	-

Tabla. 2.2: **Tamaños de las ontologías antes y después del pre-procesado**

2.2.2.1. Pre-procesado de CELDA

CELDA es un compendio de varias ontologías de interés biológico por lo que su estructura original se presenta de forma desconectada en diferentes árboles. También contiene información que no es relevante para este trabajo que debió descartarse como información relativa a tejidos, linajes celulares inmortalizados, especies, etc. Después, al tener una serie de 'árboles' de relaciones desconectados se utilizó el pre-procesado para eliminar estas discontinuidades, a continuación, se fusionaron aquellos tipos celulares que estaban repetidos tanto en humano como en ratón. Finalmente se eliminan de las etiquetas, sinónimos y nombres de las clases algunos conectores y palabras, irrelevantes para el emparejamiento por nombre, como '*the*', '*of*', '*cell*', 'cells', '*human*' y '*mouse*' además del símbolo '-'.

2.2.2.2. Pre-procesado de LifeMap

LifeMap es una ontología no está disponible en formato OWL, pero cuya información puede ser consultada en su repositorio web (Edgar et al., 2013). Para obtener la información de LifeMap en lo que respecta a desarrollo celular se lanzaron búsquedas automáticas del *website* de LifeMap que fueron recabando y construyendo el árbol de desarrollo celular, guardando información de

nombres, sinónimo y descendientes, para que sea una información comparable a la de CELDA. Esta información se almacenó en una matriz de 2 columnas en formato .ods. Después con la herramienta de pre-procesado se eliminan las palabras '*cell*', '*cells*', '*human*' y los símbolos '-', '/' y ',' de cada tipo celular.

2.2.2.3. Pre-procesado de LMHA

LMHA es una ontología específica de tipos celulares de pulmón Ardini-Poleske et al. (2017). En muchos casos no contiene información directa relacionada con el desarrollo y diferenciación celular. Por ello se utilizó el sistema de pre-procesado para eliminar clases sin información específica como '*inmunce cell*' o '*cell type*', y se añadieron nuevas relaciones y sinónimos.

En la **Tabla 2.2** se muestra el tamaño de las ontologías antes y después del pre-procesado. Estas ontologías pre-procesadas fueron las que se utilizaron para el alineamiento y fusión de ontologías.

2.2.3. Cálculo de la Matriz de mapeo por nombres

La Matriz de mapeo de nombres (*Name mapping matrix*) se construye con el objetivo de encontrar los mejores emparejamientos entre clases de dos ontologías diferentes. Para construirse puede usarse el nombre de las clases o bien el nombre junto con los sinónimos. Utilizar los sinónimos supone una mayor carga computacional, pero permite crear emparejamientos entre tipos celulares que, al tener etiquetas ambiguas, quedarían enmascarados al utilizar solo el nombre de las clases.

La Matriz de mapeo de nombres es representada mediante $S^{AB}(axb)$ donde a y b son el número de clases de las ontologías A y B respectivamente. Cada elemento $S^{AB}(ixj)$ es una medida de la similitud entre el nombre/nombre + sinónimos de cada clase i de la ontología A y cada clase j de la ontología B.

Para calcular la similitud entre dos nombres de clases se utiliza la métrica de Levenshtein (Levenshtein, 1966), la cual mide el mínimo número de inserciones, borrados y sustituciones necesarios para hacer que dos cadenas de caracteres sean iguales. Para obtener la similitud en un rango entre [0 , 1] se utiliza el opuesto a la métrica escalada de Levenshtein:

$$S^{AB}(i,j) = 1 - \frac{lev(Label_i^A, Label_j^B)}{max(|Label_i^A|, |Label_j^B|)} \qquad \text{(Ec. 2.2.1)}$$

Donde lev es la distancia de Levenshtein entre dos cadenas de caracteres. Para dos cadenas a y b de longitudes $|a|$ y $|b|$ respectivamente. La distancia de Levenshtein $lev(|a|, |b|)$ es:

$$lev\left(|a|,|b|\right) = \begin{cases} max\left(i,j\right) & if\ \min\left(i,j\right) = 0, \\ \min\begin{cases} lev\left(i-1,j\right)+1 \\ lev\left(i,j-1\right)+1 \\ lev\left(i-1,j-1\right)+1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases} \quad \text{(Ec. 2.2.2)}$$

Donde $1_{(ai \neq bj)}$ es la función indicadora igual a 0 cuando $a_i = b_j$, e igual a 1 en caso contrario, y $lev(i, j)$ es la distancia entre los primeros i caracteres de a y los primeros j caracteres de b. $Label_i^A$ y $Label_i^B$ son las etiquetas de la clase i y j de las ontologías A y B, respectivamente, y || es la longitud de la cadena. Aplicando la **Ec. 2.2.3** por pares para cada clase de etiqueta eliminada i de A, y la clase de etiqueta eliminada j de B, FOntCell construye la Matriz de mapeo de nombres S_{AB} entre A y B.

Al utilizar también los sinónimos de cada clase supone una nueva capa de cálculos, ya que no solo hay que calcular la similitud entre los nombres de cada clase si no cada uno de los sinónimos con cada sinónimo de cada clase. Aun así, la matriz S^{AB} resultante tiene la misma forma y debe darnos la misma información para el emparejamiento: Un resultado entre [0, 1] para cada clase de la ontología A contra cada clase de la ontología B.

Al utilizar los sinónimos de las clases, la similitud entre dos clases i, j se calcula utilizando una lista de sinónimos (que incluyen el nombre), con $\{i\} \in A$ y $\{j\} \in B$. Con esta lista se calcula una matriz de emparejamiento por nombres $S^{\{i\}\{j\}}(|\{i\}||\{j\}|)$ entre cada sinónimo de la lista de la clase $\{i\}$ y cada sinónimo de la lista de la clase $\{j\}$ mediante la distancia de Levenshtein (**Ec. 2.2.3**), $|\{i\}|$ y $|\{j\}|$ son el tamaño de las listas $\{i\}$ y $\{j\}$ respectivamente. Para finalizar, la mayor puntuación de $S^{\{i\}\{j\}}$ será la puntuación de alineamiento elegida para la Matriz de mapeo por nombres $S^{AB}(i, j) = maxS^{\{i\}\{j\}}$.

FOntCell considera que dos etiquetas tienen un emparejamiento cuando obtienen una puntuación en la **Ec. 2.2.3** mayor o igual al umbral θ_N que por defecto se establece en $0{,}85$

2.2.4. Cálculo de la Matriz de mapeo estructural

No todas las clases de una ontología son identificables como clases de la otra ontología solo por mapeo de nombres. Por ejemplo, en el caso de la fusión de CELDA y LifeMap, cuando usamos solo el mapeo por nombre aproximadamente el 60 % de las clases de CELDA son inicialmente no asignadas a LifeMap. FOntCell permite reconocer como equivalentes dos clases con diferente etiquetado (incluso en los sinónimos) que correspondan al mismo concepto en función de la estructura de la propia ontología alrededor de estas dos clases, el mapeo estructural.

Una determinada relación en una ontología, y sus clases, puede considerarse como un grafo y sus nodos respectivamente. El Mapeo Estructural se basa en esta consideración para comparar dos clases como dos nodos de dos grafos.

Para relacionar dos nodos de dos ontologías, FOntCell extrae un subgrafo local centrado en cada nodo, a los cual llamaremos el nodo generador i de la ontología A y el nodo generador j de la ontología B. El conjunto de subgrafos de las ontologías A y B se designan como $\{A\}$ y $\{B\}$ respectivamente. El conjunto de nodos extraídos de los subgrafos creados por los nodos generadores i y j se designan como $\{i\}$ y $\{j\}$ respectivamente. El tamaño de estos subgrafos se establece por el parámetro W, el cual indica el número de generaciones de ascendencia y descendencia desde el nodo generador que FOntCell toma para crear el subgrafo (**Fig. 2.3**). FOntCell mide la similitud estructural entre dos grafos utilizando diferentes métodos. El objetivo es construir la Matriz de mapeo estructural *Structure Mapping Matrix* $T^{AB}(axb)$ donde a y b son el número de clases de las ontologías A y B respectivamente. Una vez se ha establecido el tamaño de la ventana W (por defecto 4), para cada nodo i de A FOntCell construye el subgrafo que le rodea con nodos $\{i\} \in A_w(i)$ y calcula su similitud con todos los nodos $\{j\} \in B_w(j)$ donde $A_w(i)$ y $B_w(j)$ son los subgrafos de tamaño W centrados en los nodos i y j respectivamente. Así, FOntCell hace un emparejamiento estructural convolucional, tejiendo diferentes métricas para calcular la similitud entre los subgrafos $A_w(i)$ y $B_(j)$.

El método de Blondel (Blondel et al., 2004) inicialmente desarrollado para medir la similitud entre los vértices de grafos, puede ser utilizado para calcular el emparejamiento estructural entre dos redes, aunque implica mucha demanda computacional (**Fig. 2.5**). Para mejorar la velocidad del mapeo estructural se diseñaron dos nuevos métodos para calcular el emparejamiento estructural de forma convolucional: el emparejamiento estructural basado en vectores y el emparejamiento de estructuras basado en restricción. También se adaptó el método de Blondel para que trabajara de forma convolucional. Un ejemplo del desplazamiento de la ventana de forma convolucional se muestra en la figura **Fig. 2.3**. FOntCel toma como nodos generadores aquellas clases que no han tenido ninguna equivalencia durante el mapeo por nombre, es decir, las clases que no estén presentes en la Matriz de mapeo por nombres S^{AB}.

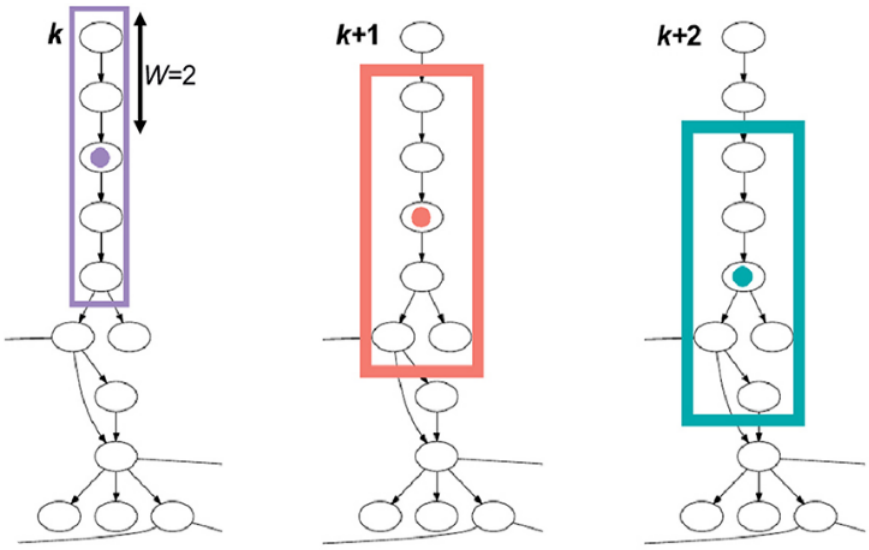


Figura 2.3: **Ejemplo de tres pasos consecutivos del deslizamiento de una ventana de tamaño $w = 2$ utilizado en el cálculo del emparejamiento convolucional por estructura.** Para cada nodo central (generador), marcado con un círculo de color, los nodos involucrados en el cálculo del emparejamiento convolucional por estructura se enmarcan con un rectángulo del mismo color que su correspondiente nodo central.

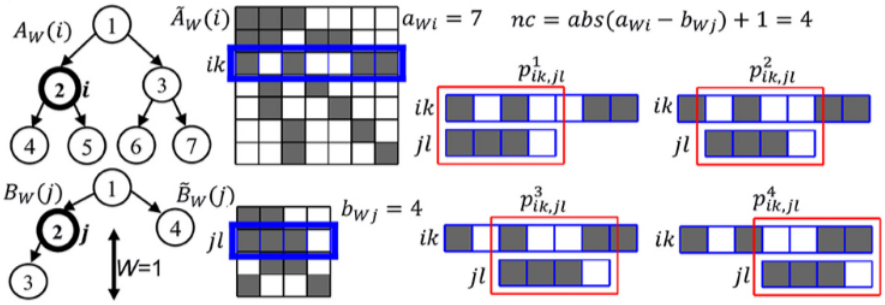


Figura 2.4: **Ejemplo de convolución de grafo.** Deslizamiento de la ventana de tamaño $w = 1$ entre dos subgrafos $A_w(i)$ y $B_w(j)$ con nodos generadores i y j (a la izquierda), sus matrices de adyacencia $\widetilde{A}_w(i)$ y $\widetilde{B}_w(j)$ (en el centro) y con número de nodos $a_{wi} = 7$ y $b_{wj} = 4$, respectivamente. Los nodos conectados son representados por celdas oscuras en las matrices de adyacencia. Para cada fila k de $\widetilde{A}_w(i)$ y l de $\widetilde{B}_w(j)$ se calcula una convolución vectorial. En azul se resalta el ejemplo del paso de $k = 3$ de $\widetilde{A}_w(i)$ y $l = 2$ de $\widetilde{B}_w(j)$. En rojo (a la derecha) se resalta el $nc = abs(a_{wi} - b_{wj}) + 1 = 4$ del deslizamiento de la ventana de la fila más corta: jl de $\widetilde{B}_w(j)$ sobre la más fila más larga: ik de $\widetilde{A}_w(i)$. Se añaden también las respectivas similitudes por convolución nc, $p^c_{ik,jl}$, donde cada fragmento c se calcula utilizando una una de las métricas $M = \{1 - coseno, Euclidean, 1 - Pearson\}$.

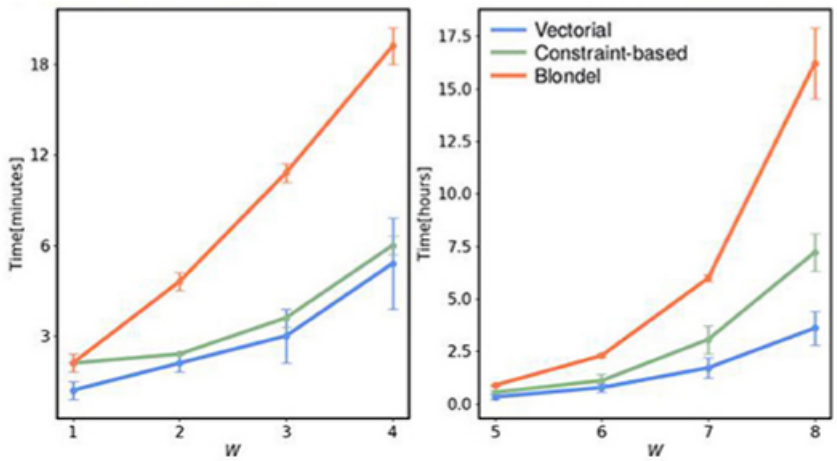


Figura 2.5: **Tiempo de ejecución de los cinco métodos de emparejamiento estructural.** Con los parámetros: $\theta_N = 0{,}85$ y $\theta_{LN} = 0{,}7$ con tamaños de ventana w en un rango de [1, 8]. El método vectorial de emparejamiento estructural {coseno, Euclídeo y Pearson} tienen tiempos de ejecución similares por lo que se representan por una misma línea.

Cada uno de los métodos desarrollados a continuación genera una puntuación estructural entre los nodos (i, j) definidos en $T^{AB}(i, j)$ donde $i \in A$ y $j \in B$. La convolución mejora el resultado del mapeo estructural en lugar de utilizar el grafo completo de la matriz para calcular los emparejamientos, ya que reduce la influencia de nodos y aristas distantes. El Mapeo por Nombre y el Mapeo por Estructura portan información complementaria.

2.2.4.1. Similitud Vectorial Basada en Convolución de Grafos como Métrica de Emparejamiento Estructural

Para cada posible par de nodos $i \in A$ y $j \in B$, y su tamaño de ventana W, se generan los subgrafos $A_w(i)$ y $B_w(j)$ centrados en i y en j, y las matrices de adyacencia $\widetilde{A}_w(i)$ y $\widetilde{B}_w(j)$ con número de nodos a_{wi} y b_{wj}, respectivamente. Para todos los posibles pares de nodos $k \in A_w(i)$ u $l \in B_w(j)$, se toman las correspondientes filas $\widetilde{ik}$ y $\widetilde{jl}$ de las matrices de adyacencia $\widetilde{A}_w(i)$ y $\widetilde{B}_w(j)$ y se calcula su similitud utilizando uno de las métricas: $M = \{1 - Coseno, Euclidean, 1 - Pearson\}$, a lo largo del texto los llamamos como emparejamiento por coseno, emparejamiento Euclídeo y emparejamiento por Pearson, respectivamente. Como los tamaños de a_{wi} y b_{wj} de esas filas no son necesariamente iguales, se calculan todos las nc posibles similitudes convolucionales, $p^c_{ik,jl}$, de la fila más corta sobre la más larga, donde $nc = abs(a_{wj} - b_{wi}) + 1$ es el número de convoluciones y $c \in [1, nc]$ (**Fig. 2.4**), y selecciona la similitud máxima $p^{Max}_{ik} = max p^{Max}_{xik,jl}$ para jl a lo largo de $\widetilde{B}_w(j)$.

2.2.4.2. Similitud Basada en Restricción como Métrica de Emparejamiento Estructural

El emparejamiento basado en restricciones (*constraint-based*) tiene tres suposiciones:

- (i) Los emparejamientos obtenidos por emparejamiento de nombre son correctos.
- (ii) El grado de similitud estructural de dos nodos generadores es proporcional al número de emparejamientos entre los subgrafos generados por ellos.
- (iii) Dos nodos generadores tienen una mayor posibilidad de ser equivalentes si sus parientes cercanos tienen emparejamientos por nombre entre ellos.

Para calcular la similitud entre dos nodos generadores, i y j, en primer lugar, se obtiene el número de emparejamientos por nombre entre los conjuntos de nodos $\{i\}$ y $\{j\}$ de los subgrafos, de tamaño w centrados en los nodos i y j, $A_w(i)$ y $B_w(j)$. A continuación, se calcula una ponderación de acuerdo con su cercanía a los nodos generadores i y j. Los emparejamientos más cercanos al nodo generador tienen una mayor puntuación.

Para implementar esté método, para todos los pares de nodos posibles $i \in A$ y $j \in B$, y para una

ventana de tamaño w, se buscan todos los posibles emparejamientos entre las listas de nodos $\{k\} \in A_w(i)$ y la lista $\{l\} \in B_w(j)$ con la condición de que al menos un nodo de la lista $\{k\} \in A_w(i)$ tenga un emparejamiento por nombre con un nodo de la lista $\{l\} \in B_w(j)$. Entonces, para cada nodo k en la lista $\{k\}$ se aplica el algoritmo de grafos del camino más corto (*shortest path*) s_{ki} a i y se produce la lista $\{s_{ki}\}$ de los caminos más cortos para estimar la cercanía de cada nodo $\{i\}$ al nodo generador i. Se designa a cada s_{ki} una restricción (*constraint*) $c_{ki} = W + 1 - s_{ki}$. Finamente se suman la lista de restricciones $\{c_{ki}\}$ para crear la restricción acumulada C_i que es la que se asigna en la Matriz de mapeo estructural $T^{AB}(i, j)$.

2.2.4.3. Similitud de Blondel como métrica de emparejamiento estructural

Para calcular el emparejamiento estructural, se adaptó la métrica original de Blondel:

$$T_{k+1}^{AB} = \frac{\widetilde{B} T_k^{AB} \widetilde{A}^t + \widetilde{B}^t T_k^{AB} \widetilde{A}}{||\widetilde{B} T_k^{AB} \widetilde{A}^t + \widetilde{B}^t T_k^{AB} \widetilde{A}||} \qquad \text{(Ec. 2.2.3)}$$

Donde t es el operador de transposición. La **Ec. 2.2.4.3** se calcula iterativamente hasta que se obtiene un número par de pasos k hasta que se alcanza la convergencia estable del matriz emparejamiento de estructura T^{AB}. Como en las otras métricas se toman subgrafos generados de cada nodo y se calcula la similitud entre esos subgrafos utilizando la métrica de Blondel. Para cada nodo i de A se construye un subgrafo $\{i\} \in A$ y se calcula su similitud con el subgrafo $\{j\} \in A$ utilizando la **Ec. 2.2.4.3** con las matrices de adyacencia de cada subgrafo. Se genera una estructura de convolución tejiendo la **Ec. 2.2.4.3** para todos los subgrafos generados a partir de $\{i\}$ y $\{j\}$.

$$T_{k+1}^{\{i\}\{j\}} = \frac{\widetilde{\{j\}} T_k^{\{i\}\{j\}} \widetilde{\{i\}}^t + \widetilde{\{j\}}^t T_k^{\{i\}\{j\}} \widetilde{\{i\}}}{||\widetilde{\{j\}} T_k^{\{i\}\{j\}} \widetilde{\{i\}}^t + \widetilde{\{j\}}^t T_k^{\{i\}\{j\}} \widetilde{\{i\}}||} \qquad \text{(Ec. 2.2.4)}$$

Donde $\widetilde{\{i\}}$ y $\widetilde{\{j\}}$ son el conjunto de matrices de adyacencia de los respectivos subgrafos generados de $\{i\}$ y $\{j\}$. Finalmente, la puntuación estructural en las posiciones i y j en la matriz $T_{k+1}^{\{i\}\{j\}}$ es asignada a la Matriz de mapeo estructural $T^{AB}(i, j)$.

2.2.5. Emparejamiento de ontologías por nombres locales

Para mejorar el resultado obtenido con el método de emparejamiento estructural se realiza una comparación adicional de nombres locales utilizando la Matriz de mapeo por nombres S^{AB} para calcular la media del emparejamiento por nombre $S^{\{i\}\{j\}}$ de cada par de nodos $\{i\}\{j\}$, con el mismo tamaño de ventana w utilizado para calcular $T^{\{i\}\{j\}}$. Se seleccionan las mejores puntuaciones del emparejamiento de nombre de $S^{AB}(i, j)$, se calcula la media de esta puntuación de emparejamiento por nombre para el par $\{i\}\{j\}$ y se construye una nueva matriz de emparejamiento por

nombre de $\{i\}\{j\}$: $S^{\{i\}\{j\}}$. Si $S^{\{i\}\{j\}} > \Theta_{LN}$ donde Θ_{LN} es el umbral para el emparejamiento por Nombres locales (por defecto $\Theta_{LN} = 0{,}7$), se consideran los nodos i y j como emparejamiento estructural.

2.2.6. Alineamiento de ontologías

Para emparejar clases, FOntCell inicialmente elige el mejor emparejamiento para cada nodo i de la ontología A con un nodo j de la ontología B, utilizando la Matriz de mapeo por nombres S^{AB}. Si $S^{AB}(i,j) > \Theta_N$, (siendo Θ_N el mapeo por nombre elegido por el usuario), se considera que las clases i y j como emparejadas y se clasifica esta asignación como 'name match' (**Fig. 2.2**). Si $S^{AB}(i,j) \leq \Theta_N$ se toma el elemento $T^{AB}(i,j)$ de uno de los métodos de emparejamiento estructural elegidos por el usuario a la hora de utilizar FOntCell. De esta forma se obtiene el mapeo estructural, y se considera un emparejamiento estructural cuando $T^{AB}(i,j) \geq \Theta_T$ siendo Θ_T el umbral de mapeo estructural elegido por el usuario (**Fig. 2.2**).

Como resultado del alineamiento se crea un documento con la información relevante de cada nodo de la ontología A. El documento cuenta con 5 columnas:

1. Etiquetas de las clases de la ontología A
2. Etiquetas asignadas de las clases de B a las clases de A
3. Puntuación por emparejamiento de nombre
4. Puntuación por emparejamiento por estructura
5. Tipo de emparejamiento (*Name*/*Structure*). En caso de no existir una asignación el tipo de asignación se señala como *Non-matched*.

2.2.7. Fusión de ontologías

Una vez que las clases entre las dos ontologías han sido emparejadas se procede a 'traducir' los nombres/etiquetas de todas las clases de la ontología B a sus nombres equivalentes, si los hay, de la ontología A. Después se añaden a la ontología A toda la descendencia que tuvieran las clases de B que han sido traducidas, añadiéndolas a sus equivalentes clases de A. A continuación, se eliminan aquellas posibles repeticiones de relaciones entre clases que se hayan podido generar. El *array* resultante de relaciones es la mezcla de las dos ontologías. Además, se genera un documento en formato OWL con el resultado de la fusión, el cual es como el documento .owl de la ontología A al cual se le añaden todas las nuevas clases y relaciones que se extraen de la

ontología B teniendo en cuenta las equivalencias. A estas nuevas clases se les otorga en la ontología resultante: una nueva ID, una etiqueta de la clase, sinónimos de la clase y una relación de ascendiente directo.

Finalmente, se crea un documento .html con los grafos acíclicos dirigidos (*Directed Acyclic Graph* (DAG)) de las ontologías originales y de la ontología fusionada, además de información estadística sobre la fusión, como puede ser: porcentaje y número de clases/relaciones añadidas y tipos de emparejamientos, de forma gráfica y textual.

2.2.8. Cálculo la puntuación de los alineamientos

Para poder evaluar el emparejamiento de FOntCell y compararlo con otros métodos de emparejamientos se utilizan las métricas de Precisión (**Ec. 2.2.8**), Recuerdo (*Recall*) (**Ec. 2.2.8**) y Exactitud (**Ec. 2.2.8**) en términos de errores de Tipo I y Tipo II:

$$Precisión = \frac{TP}{TP + FP} \quad \text{(Ec. 2.2.5)}$$

$$Recuerdo = \frac{TP}{TP + FN} \quad \text{(Ec. 2.2.6)}$$

$$F_\beta = \frac{(1+\beta^2)\cdot Precisión \cdot Recuerdo}{(\beta^2 \cdot Precisión)+Recuedo} = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP+\beta^2 \cdot FN+FP} \quad \text{(Ec. 2.2.7)}$$

Donde β es un número real positivo que da cuenta de cuántas veces la Precisión se considera más importante que el Recuerdo en la medición del Exactitud. TP, FP y FN son los números Verdaderos Positivos (*True Positives*), Falsos Positivos (*False Positives*) y Falsos Negativos (*False Negatives*) respectivamente. Se calcularon tres Exactitudes: F_1, media armónica de la Precisión y el Recuerdo, $F_{0,5}$ el cual da el doble de peso a la Precisión en comparación con el Recuerdo, atenuando la influencia de los FN y F_2 el cual da el doble del peso al Recuerdo comparado con la Precisión, dando más énfasis a los FN.

Como el emparejamiento de los casos de CELDA + LifeMap y CELDA + LifeMap + LMHA son emparejamientos *de novo*, es decir, no existen referencias de los mismos, hemos construido manualmente las referencias utilizadas para compararlas y poder así evaluar el resultado de todas las herramientas de emparejamiento que incorpora FOntCell.

2.3. Resultados

2.3.1. Búsqueda de parámetros óptimos para la fusión de CELDA y LifeMap

Para encontrar los parámetros óptimos de FOntCell para la fusión de CELDA y LifeMap. Se realizó un análisis bidimensional de los parámetros de alineamiento: umbral de Nombres locales θ_{LN} y tamaño de la ventana w en los rangos [0.1, 0.8] y [1, 8] respectivamente, utilizando variaciones de 0.1 para θ_{LN} y de 1 para w para todas las métricas de mapeo estructural: las tres métricas vectoriales (Euclídea, coseno y Pearson), la métrica de alineamiento basada en restricción y la métrica basada en el método de Blondel (**Fig. 2.6**). El método basado en restricción no utiliza el alineamiento de Nombres locales, por lo que θ_{LN} no se utiliza.

Un umbral de $\theta_N = 0{,}85$ produce un $F_1 > 0{,}9$ para todas las métricas (**Tabla 2.4**). Por lo que se mantiene el $\theta_N = 0{,}85$ para el resto del análisis. $\theta_N > \Theta_{LN}$ ayuda a recuperar muchos casos relevantes durante el mapeo estructural y también, ayuda con el problema de isomorfismo de grafos que surge durante las comparaciones de los subgrafos. El umbral del mapeo por nombre $\theta_N = 0{,}85$ asigna como clases equivalentes aquellas que tienen etiquetas con diferentes variantes ortográficas, como terminaciones en 's', apostrofes, etc. Por lo tanto, se estableció para el resto del análisis que $\theta_N > \Theta_{LN} = 0{,}7$ ya que se espera mucha más variabilidad en los nombres de los nodos al comparar subgrafos enteros que entre comparaciones directas entre clases. Es necesario reducir la sensibilidad de θ_{LN} ya que una alta sensibilidad, es decir, un alto umbral nos llevaría a solo encontrar isomorfismos en aquellas regiones que ya han sido alineadas mediante el mapeo por nombre.

Pequeños tamaños de ventana w producen subgrafos menores, esto aumenta la posibilidad de en las métricas de emparejamiento estructural encuentren mayor isomorfismo, mientras que un tamaño de ventana w muy grande provoca que el umbral θ_{LN} pierda eficacia, creando emparejamientos falsos positivos **Fig. 2.6**.

Para la fusión de CELDA + LifeMap, el tamaño de la ventana que minimizaba la sensibilidad de θ_{LN} era $w = 4$. Del análisis de la relación entre θ_{LN} y w para los diferentes métodos de mapeo estructural, que se muestran en la **Fig. 2.6**, podemos deducir:

- El método basado en restricción con una ventana $w = 4$ adolece en el problema del isomorfismo de grafos.
- El método Euclídeo es más restrictivo que los otros métodos vectoriales, pero mucho más sensible a los cambios de tamaño de θ_{LN} que los otros métodos.
- El método de Pearson, el de Blondel y el de coseno se comportan de forma similar a la

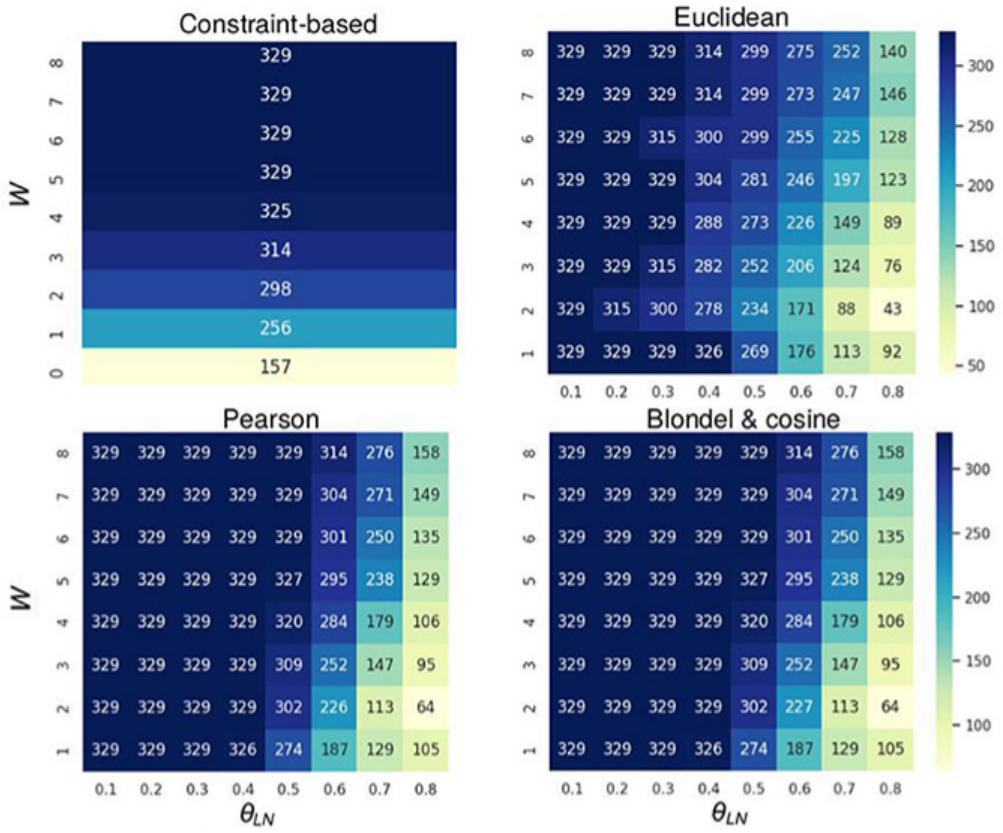


Figura 2.6: **Mapas de color de los emparejamientos por *structure mapping*.**. Obtenidos mediante combinaciones de dos parámetros: tamaño de la ventana y umbral de emparejamientos por nombre local, utilizando los cinco métodos de emparejamiento estructural: los tres métodos de emparejamiento vectorial {coseno, Euclídeo, Pearson}, emparejamiento basado en restricción (*constraint-based*) y emparejamiento por el método Blondel. los dos parámetros optimizados son el tamaño de la ventana w y el umbral de nombre local θ_{LN} en los rangos [1, 8] y [0.1, 0.8], respectivamente. Utilizando pasos de 0.1 para θ_{LN} y 1 para w. Tonos de azul más oscuros representan un mayor número de equivalencias.

variación de estos dos parámetros ya que obtienen unos números de emparejamientos similares.

Para analizar el efecto sobre el porcentaje de emparejamientos, nuevas relaciones y nuevas clases de cada uno de los cinco métodos de mapeo estructural se realizó una fusión de CELDA y LifeMap, con los parámetros optimizados $w = 4, \theta = 0{,}85$ y $\theta_{LN} = 0{,}7$, para cada método de mapeo estructural. Se encontró un número similar en el porcentaje de clases y relaciones añadidas, así como en el porcentaje de número de emparejamientos (**Fig. 2.7**). Estudiamos el tiempo de ejecución de los cinco métodos de mapeo estructural para los parámetros optimizados $\theta = 0{,}85$ y $\theta_{LN} = 0{,}7$ y un tamaño de ventana w en un rango [1, 8] y encontramos que los métodos vectoriales (coseno, Euclídeo y Pearson) son los más rápidos, incluso con órdenes de magnitud más

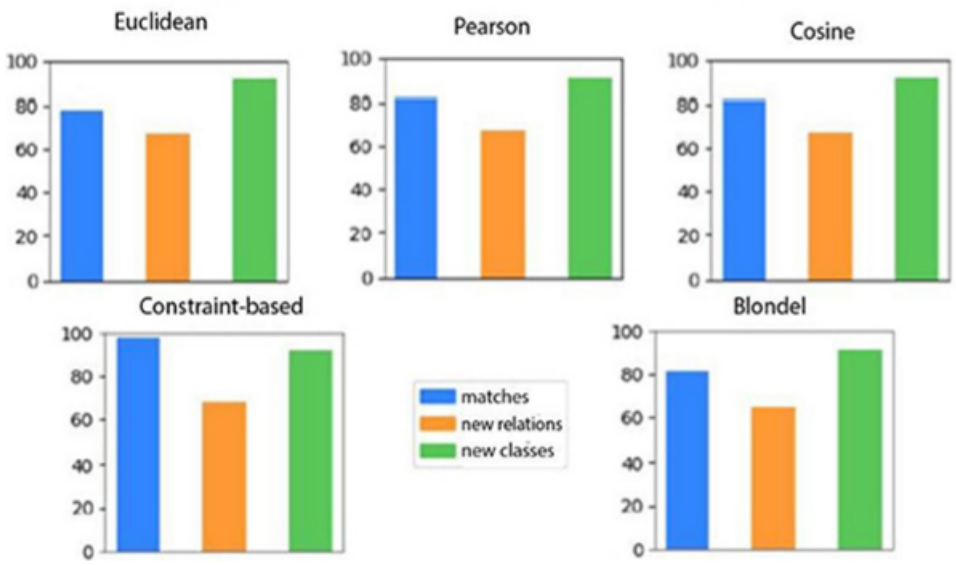


Figura 2.7: **Porcentaje de emparejamientos, nuevas clases y nuevas relaciones**. Obtenido con los cinco métodos de emparejamiento estructural y con los parámetros de alineamiento $w = 4$, $\theta_N = 0{,}85$ y $\theta_{LN} = 0{,}7$.

rápidos que el método Blondel (**Fig. 2.5**).

2.3.2. Alineamiento y fusión de CELDA y LifeMap

La fusión de CELDA y LifeMap con $\theta_{LN} = 0{,}7$ y $w = 4$ genera una ontología que integra las 841 clases de CELDA con 567 clases de LifeMap: La información celular de la ontología CELDA se incrementa un 67 % (**Fig. 2.8**) con un Exactitud $F_1 = 0{,}9$ (**Tabla 2.4**). DAG de CELDA, LifeMap y la ontología resultante se muestran en la **Fig. 2.9**.

En la **Fig. 2.10** se muestra el zoom realizado a una determinada región donde se aprecia el resultado de un emparejamiento y fusión. Esta región representa una zona equivalente en ambas ontologías, ya que ambas empiezan por el mismo tipo celular emparejado mediante emparejamiento por nombre y puede observarse que todas las células que intervienen son equivalentes.

En este zoom podemos ver como el emparejamiento estructural 'rescata' información de una ontología para aumentar la ontología fusionada final. Dos o más clases de CELDA pueden alinearse con una clase de LifeMap, un fenómeno más común si activamos el uso de sinónimos. Los subgrafos mostrados en A (CELDA) y B (LifeMap) de la **Fig. 2.10** tienen una forma similar pero no idéntica: CELDA empieza con "hypoblast cell," con el descendiente "yolk cell" y "extraembryonic endoblast cell" y con "secondary yolk sac" como descendiente de estos últimos, sin embargo, en LifeMap tenemos una línea de diferenciación con la secuencia: "hypoblast cell", "extraembryonic endoderm cells", "yolk sac endoderm cells" y finalmente "allantois cell."

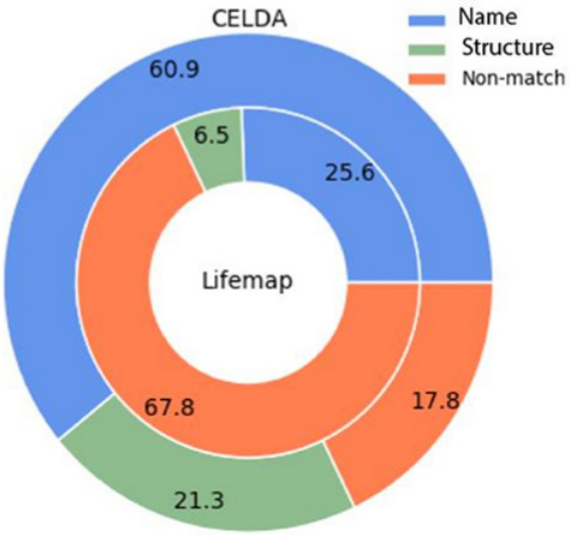


Figura 2.8: ***Donut plot*** de los porcentajes de clases añadidas mediante mapeo por nombre *vs.* las clases añadidas mediante mapeo estructural, para CELDA (círculo externo) y LifeMap (Circulo interno)

El consenso de la fusión (**Fig. 2.10C**) empieza con “hypoblast cell”, sigue con “yolk cell” e incorpora como descendiente adicional “extraembryonic endoblast cell” manteniendo la información de CELDA, a continuación, como descendiente de “yolk cell” y “extraembryonic endoblast cell” se añade “secondary yolk” como equivalente de zolk sac endoderm cells"de LifeMap, para finalmente añadir .allantois cellçomo información adicional de LifeMap. Para el par de parámetros menos restrictivo $\theta_{LN} = 0{,}1 yw = 1$ un 39 % de las clases de CELDA tuvieron un emparejamiento estructural en LifeMap independientemente del método de emparejamiento estructural utilizado. Utilizando parámetros más restrictivos $\theta_{LN} = 0{,}7 yw = 7$ se obtiene un 32 % de clases con el método estructural.

2.3.3. Estimación de la Precisión, Recuerdo y Exactitud de la fusión de CELDA y LifeMap

Se calculó la Precisión de los diferentes métodos de mapeo de FOntCell durante la fusión de CELDA y LifeMap con los parámetros óptimos $w = 4, \theta_{LN} = 0{,}7 y \theta_N = 0{,}85$ (**Fig. 2.11**). Los resultados obtenidos se validaron buscando fallos, falsos positivos (FP) y éxitos, verdaderos positivos (TP) en el emparejamiento de clases y calculando la precisión utilizando la **Ec. 2.2.8**. El emparejamiento por nombre muestra la más alta precisión del 98.63 % (**Fig. 2.12b**) y tiene el mayor número de emparejamientos (**Fig. 2.10**), 512. Entre los métodos de mapeo por estructura la mayor Precisión (62.10 %) la obtiene el método basado en restricción, seguido del método de coseno y el de Pearson56.42 %, Blondel 50.27 % y Euclídeo 48,99 % (**Fig. 2.12a**).

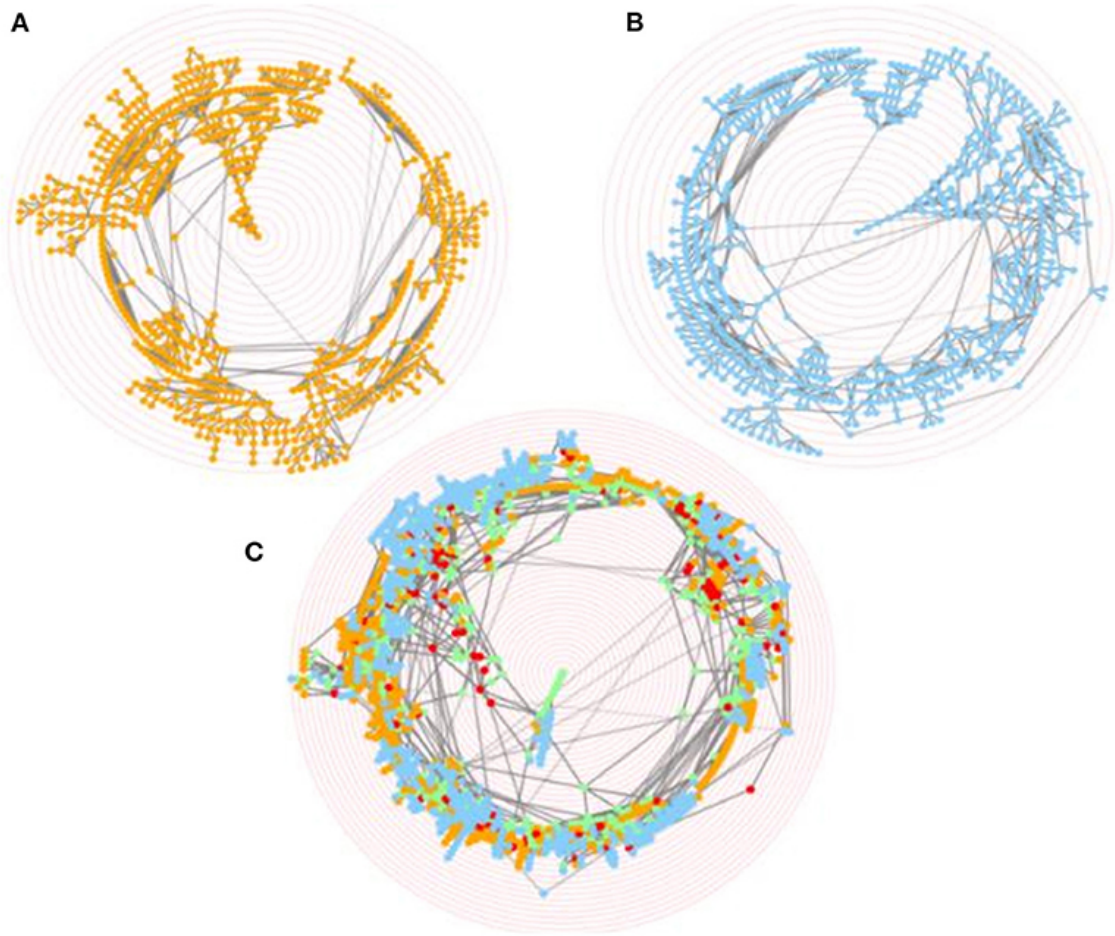


Figura 2.9: **Grafos dirigidos acíclicos (*Directed Acyclic Graphs*, DAGs)** de (A) CELDA, (B) LifeMap y (C) la fusión de CELDA + LifeMap. Los nodos azules y naranjas son aquellos no emparejados de la ontología A y B respectivamente. Los nodos verdes y rojos son aquellos emparejados por nombre o por estructura respectivamente. Las etiquetas de ontología. Los anillos rojos concéntricos son guías para el zoom.

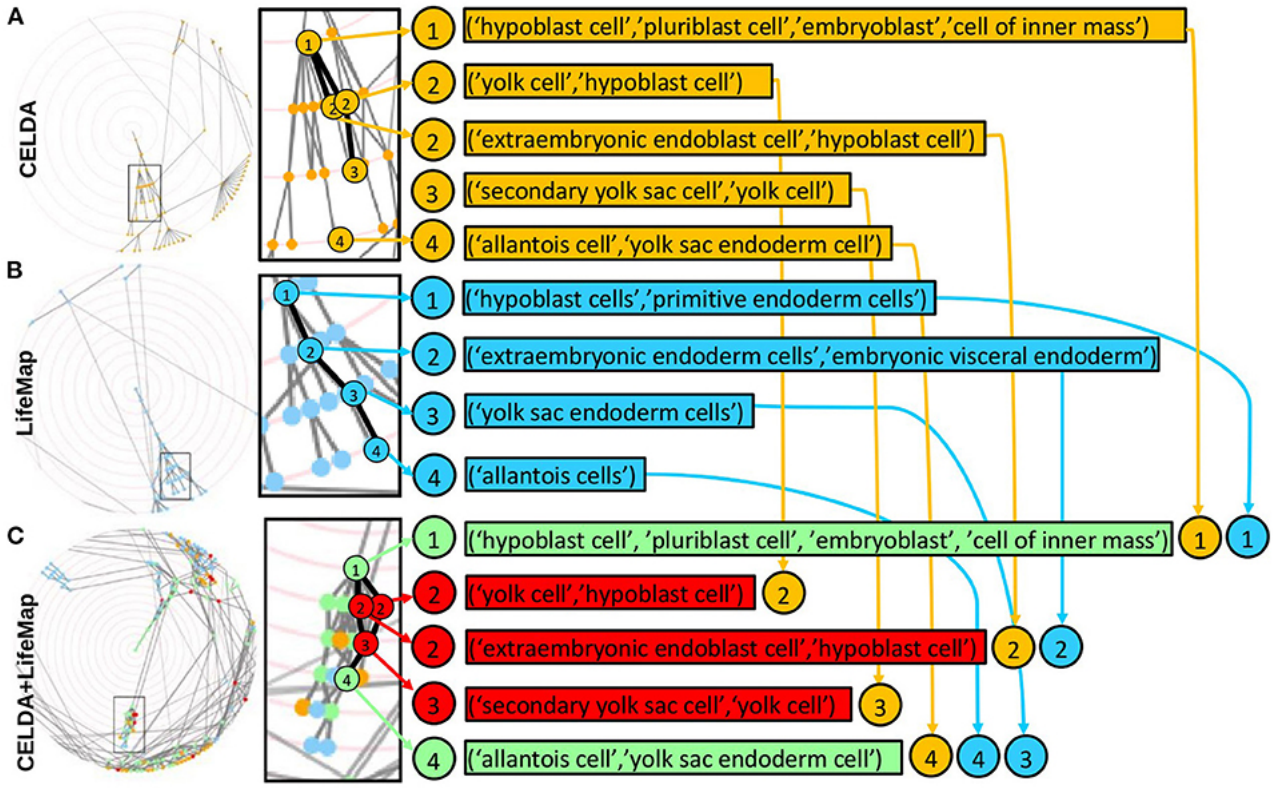


Figura 2.10: **Zooms en regiones de CELDA, LifeMap y la ontología resultante de la fusión.** En este *zoom* se observa el mapeo estructural y de nombre. A la izquierda se muestra la captura de pantalla de los Grafos Dirigidos Acíclicos de CELDA (A), LifeMap (B) y la fusión CELDA+LifeMap (C). A la derecha se muestra el zoom de las regiones con los correspondientes tipos celulares, de los cuales se muestran sus sinónimos. Los nodos naranjas y azules son aquellos que hacen referencia a CELDA y LifeMap, respectivamente, y en la ontología resultante son aquellos que no han tenido ningún emparejamiento. Los verdes y rojos son aquellos que sí han tenido un emparejamiento, ya sea mediante mapeo por nombre o por estructura, respectivamente. Los números dentro de los círculos indican la generación, es decir el grado de proximidad al nodo padre, en orden ascendente.

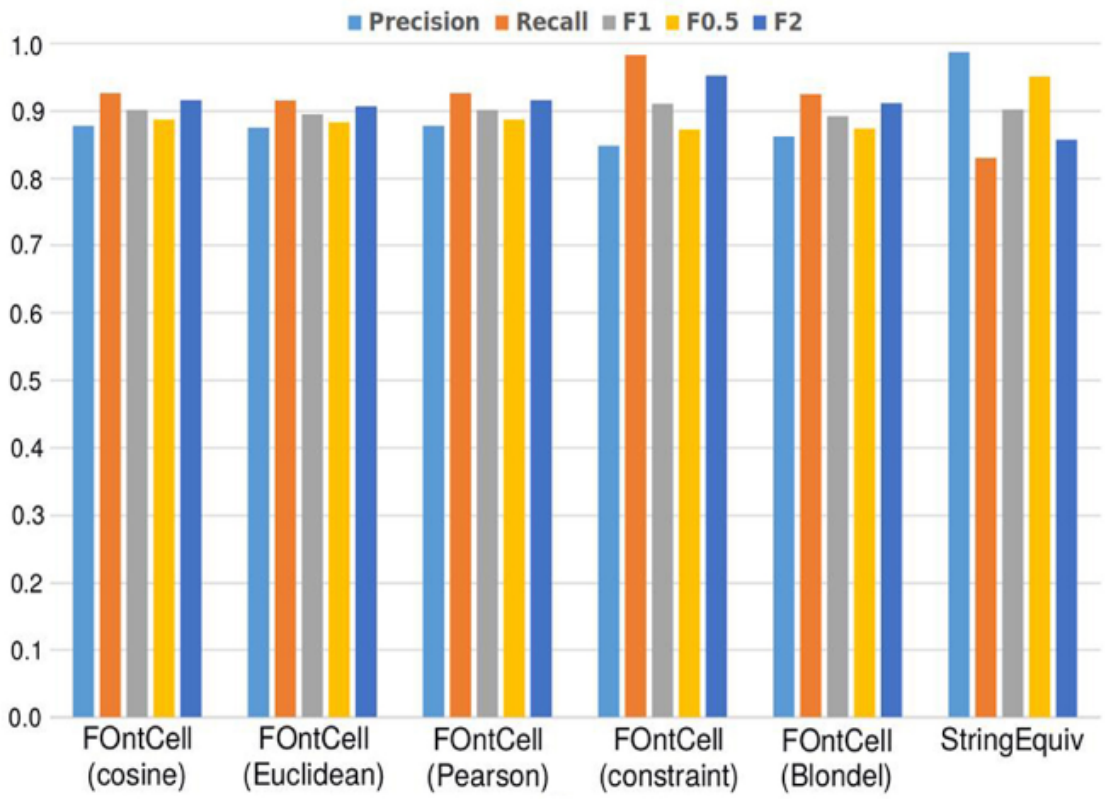
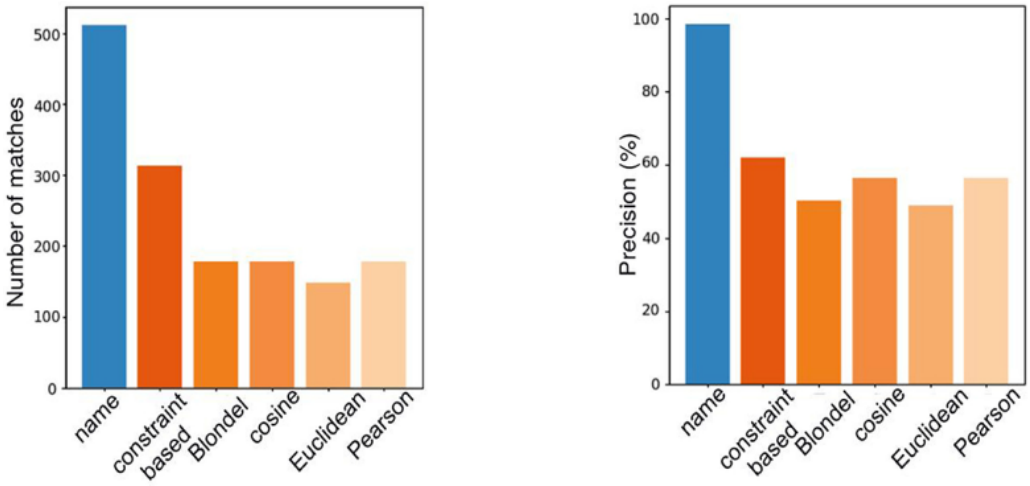


Figura 2.11: **Precisión, Recuerdo y F_β de las diferentes métricas de alineamiento mediante mapeo estructural.** Combinado con el mapeo por nombre (FOntCell) y el mapeo por nombre aplicado por separado (StringEquiv). Los parámetros de alineamiento son: $W = 4$ $\theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$

Para evaluar todo el proceso de mapeo de FOntCell, se tuvieron en cuenta ambos métodos de mapeo, por nombre y por estructura. Se observó una Precisión similar entre todos los métodos ∼87 % utilizando los métodos vectoriales, 86.1 % con el método de Blondel, y 86 % con el método basado en restricción. Los métodos vectoriales producen valores mayores de Precisión principalmente debido a un menor número de emparejamientos que en el método por restricción y en el método de Blondel.

Cuando se considera solo la Precisión del mapeo estructural, el método de Blondel es el segundo peor, ligeramente mejor que el método Euclídeo (**Fig. 2.12b**). A pesar de ello, cuando se combina con el emparejamiento por Nombres locales, todos los métodos vectoriales, incluido el método Euclídeo superan la precisión del método de Blondel (**Fig. 2.11**) debido a que el método Blondel produce más emparejamientos durante el emparejamiento estructural comparado con el Euclídeo (**Fig. 2.12a**). Esto indica que existe una sinergia entre el emparejamiento por Nombres locales y el mapeo estructural y que la combinación es más fuerte en el caso de los métodos vectoriales que en el caso de Blondel, al menos para la fusión de CELDA y LifeMap. El método basado en restricción tiene la Precisión más alta entre todos los métodos de mapeo estructural (**Fig. 2.12b**)



(a) (b)

Figura 2.12: **diagramas de barras del alineamiento de ontologías.** (a) Número de emparejamientos durante el alineamiento de ontologías con diferentes métodos de mapeo. El mapeo por nombre se muestra en azul y el estructural en diferentes tonos de naranja. Los parámetros de alineamiento son: $W = 4\ \theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$ (b) Precisión del emparejamiento por nombre y por los diferentes métodos de mapeo estructural. El mapeo por nombre se muestra en azul y el estructural en diferentes tonos de naranja. Los parámetros de alineamiento son: $W = 4\ \theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$

y tiene una Precisión total inferior una vez se combina con el emparejamiento por nombres locales (**Fig. 2.11**).

Los métodos Pearson y coseno muestran unos resultados similares, ambos tienen el mismo número de emparejamientos, 179 (**Fig. 2.12a**), y la misma Precisión en el mapeo estructural (56.42 % **Fig. 2.12b**), con resultados de Precisión de 87.69 % cuando se combina con el emparejamiento por Nombres locales y el mapeo por nombres (**Fig. 2.11**). En conclusión, los métodos de coseno y Pearson en combinación con el mapeo por nombres alcanzan la mayor Precisión y el menor número de emparejamientos.

Además de la Precisión (**Ec. 2.2.8**, el Recuerdo (**Ec. 2.2.8**) y la familia de Exactitud es $F_\beta : F_1, F_{0,5} y F_2$ (**Ec. 2.2.8**) con los parámetros óptimos: $W = 4\ \theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$ para cada métrica estructural (coseno, Euclídea, Pearson, basada en restricción y Blondel) y la métrica adicional que solo utiliza alineamiento por nombre.

Todos los métodos generan alineamientos de CELDA y LifeMap con F_β parecidos. Coseno y Pearson obtienen valores F_1 y F_2 ligeramente superiores que el Euclídeo debido al bajo Recuerdo que obtiene el método Euclídeo. El método Blondel muestra un resultado similar a los métodos vectoriales con un ligero descenso de la Precisión y un Recuerdo similar al del coseno y Pearson. El

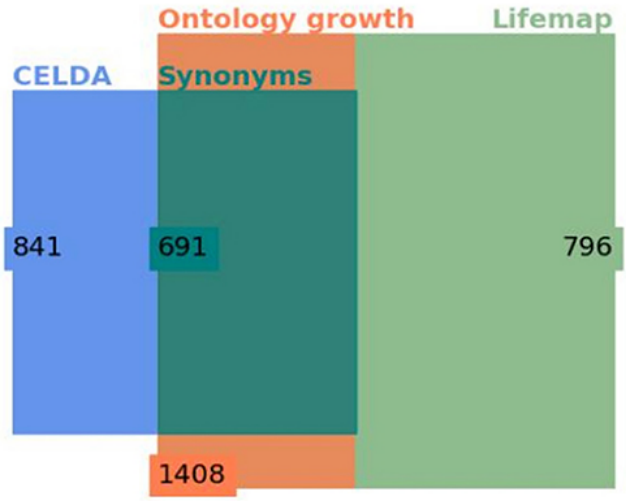


Figura 2.13: **Diagrama de cuadrados 'Euler-Venn' con el número de clases antes y después de la fusión.** En azul y verde claro se muestran los rectángulos con las clases de CELDA y LifeMap, respectivamente, antes de la fusión. El cuadrado verde oscuro muestra la suma de equivalencias tantos por nombre como por estructura. El rectángulo naranja muestra el número total de clases en la ontología resultante de CELDA + LifeMap. Los parámetros de alineamiento son: $W = 4$ $\theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$

método basado en restricción tiene la menor Precisión, pero el mayor Recuerdo (**Tabla 2.4**).

Para todos los métodos de alineamiento, los rangos de Precisión oscilan entre 0.847 y 0.877, el Recuerdo entre 0.982 y 0.915 (**Tabla 2.4**). En el caso de utilizar únicamente el mapeo por nombre (StringEquiv), la Precisión se acerca a 1, pero el Recuerdo desciende considerablemente (**Tabla 2.4**), esto es debido a que el mapeo por nombre deja sin emparejar numerosas clases que puede emparejar el método de emparejamiento estructural.

Después de la fusión de las ontologías alineadas, las ontologías resultantes del método Pearson y coseno tienen un mayor número de emparejamientos y un mayor crecimiento en el número de clases (**Fig. 2.5**) comparado con el método Euclídeo. El parámetro que influencia más el proceso en términos de crecimiento del número de clases y número de emparejamientos en la ontología final son w y θ_{LN}. Para todos los valores de θ_N, el número de emparejamientos estructurales tienen un punto de inflexión cuando $w = 4$ y $\theta_{LN} = 0{,}7$ (**Fig. 2.6**), este es el punto medio donde el método no es suficientemente restrictivo y no es excesivamente permisivo.

La fusión de CELDA Y LifeMap con los parámetros óptimos: $W = 4$ $\theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$ y con el método vectorial coseno encontramos 691 sinónimos entre las dos ontologías y genera una ontología celular final con 1408 clases, 841 de CELDA y 567 clases añadidas de LifeMap (**Fig. 2.13**).

	Precision	**Recall**	F_1	$F_{0,5}$	F_2
FOntCell (coseno)	0.861	0.720	0.784	0.829	0.744
FOntCell (Euclideo)	0.909	0.726	0.807	0.865	0.756
FOntCell (Pearson)	0.859	0.724	0.786	0.828	0.748
FOntCell (Restricción)	0.846	0.718	0.777	0.817	0.740
FOntCell (Blondel)	0.860	0.724	0.786	0.829	0.748
StringEquiv	0.997	0.662	0.766	0.890	0.673
AML	0.950	0.936	0.943	0.947	0.939
LogMap	0.918	0.846	0.881	0.903	0.859
AGM	0.152	0.195	0.171	0.159	0.185
ALIN	0.974	0.698	0.813	0.903	0.740
DOME	0.996	0.615	0.760	0.886	0.666
FCAMap-KG	0.873	0.631	0.773	0.893	0.681
Lily	0.873	0.796	0.833	0.856	0.810
LogMapBio	0.872	0.925	0.898	0.882	0.914
LogMapLite	0.996	0.728	0.829	0.904	0.765
POMAP++	0.919	0.877	0.898	0.910	0.885
SANOM	0.888	0.844	0.865	0.879	0.852
Media	**0.874**	**0.727**	**0.786**	**0.834**	**0.747**

Tabla. 2.3: **Tabla comparativa entre FOntCell y las herramientas del OAEI para el problema planteado por OAEI 2019**. En el caso de FOntCell, el nombre del método de mapeo estructural está entre paréntesis. Se resalta la media en los resultados obtenidos de las otras herramientas con las que se compara FOntCell.

2.3.4. Comparación de FOntCell con otras herramientas de OAEI

Para comparar la capacidad de alineamiento de FOntCell con el resto de herramientas se comparó en primer lugar con las ontologías de anatomía de ratón y humano comparadas en el marco de la tarea propuesta para el *Ontology Alignment Evaluation Initiative* (OAEI) en 2019. Se utilizó los parámetros optimizados: $W = 4$ $\theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$, y se hizo un alineamiento para cada métrica estructural implementada en FOntCell. Para todas las métricas FOntCell obtiene resultados en torno a la media de Exactitud F_β, Precisión y Recuerdo (**Tabla 2.3**).

Para continuar con la comparativa se seleccionaron las herramientas que consiguieron los mejores resultados para el alineamiento de las ontologías de humano y ratón para la OAEI (**Tabla 2.3**: StringEquiv, AML, LogMap y se corrieron con sus parámetros por defecto para comparar su capacidad de alineamiento con FOntCell para el caso de CELDA y LifeMap. Estas herramientas mostraron una mayor Precisión, pero un menor Recuerdo comparando con FOntCell. Este bajo Recuerdo les penalizó el Exactitud obteniendo valores de F bajos (**Tabla 2.4**).

La Exactitud F_1 de StringEquiv y diferentes métodos de alineamiento de FOntCell son similares, en torno a 0.9. Para la Exactitud que da más peso a la Precisión $F_{0,5}$, StringEquiv supera al resto de métodos. Para la Exactitud que da más importancia al Recuerdo, F_2, FOntCell obtiene mejores resultados. En nuestro caso, al buscar la fusión orientada a complementar dos ontologías el Recuerdo es clave para el 'rescate' de cuantos más tipos celulares podamos. Todos los métodos de FOntCell superan a las mejores herramientas de la OAEI en términos de Exactitudes F_β y para

	Precision	**Recall**	F_1	$F_{0,5}$	F_2
FOntCell (coseno)	0.877	0.925	0.900	0.886	0.915
FOntCell (Euclideo)	0.874	0.915	0.894	0.882	0.906
FOntCell (Pearson)	0.877	0.925	0.900	0.886	0.915
FOntCell (Restricción)	0.847	0.982	0.910	0.871	0.952
FOntCell (Blondel)	0.861	0.924	0.891	0.873	0.911
StringEquiv	0.986	0.829	0.901	0.950	0.857
AML	0.971	0.269	0.422	0.639	0.315
LogMap	0.983	0.317	0.480	0.692	0.367
Media	**0.910**	**0.761**	**0.787**	**0.835**	**0.767**

Tabla. 2.4: **Tabla comparativa entre FOntCell y las herramientas del OAEI para la fusión de CELDA y LifeMap.** En el caso de FOntCell, el nombre del método de mapeo estructural está entre paréntesis. Se resalta la media en los resultados obtenidos de las otras herramientas con las que se compara FOntCell.

el alineamiento de CELDA y LifeMap (**Tabla 2.4**).

2.3.5. Fusión de CELDA + LifeMap con LMHA

Una de las aplicaciones de FOntCell es la de fusionar una ontología más generalista con una más específica dentro del mismo ámbito del conocimiento. Para ilustrar esta funcionalidad se fusionó la ontología resultante de CELDA + LifeMap con LMHA, una ontología específica de desarrollo de tipos celulares que empieza con aproximadamente 36 semanas de gestación fetal y continua después del nacimiento con alguna variación cuando comienza el estadio alveolar y cuando se completa. El documento .owl utilizado es el que ha sido generado por (Ardini-Poleske et al., 2017). La fusión de CELDA + LifeMap con LMHA generó 65 nuevas relaciones y 39 nuevas clases relacionadas con células del endotelio y linfoides (**Fig. 2.14**).

2.4. Discusión

El descubrimiento de nuevos tipos celulares como los producidos por el consorcio del HCA o su mejor caracterización por transcriptómica de célula única (Gerovska, Araúzo-Bravo, 2016) pueden dejar obsoletas algunas ontologías de desarrollo celular. Hemos desarrollado FOntCell para acatar este problema con un algoritmo que mediante la fusión de ontologías añade nuevas relaciones y clases a una ontología base, por lo que podemos construir una nueva ontología basada en tipos celulares y mantenerla actualizada con nuevas adiciones de ontologías de menor tamaño y más específicas.

Hemos implementado FOntCell como una herramienta que fusiona dos ontologías de un mismo o similar dominio del conocimiento. Para esta fusión integra un motor de búsqueda de similitud por nombre y otro de similitud estructural basado en la convolución de grafos. El cálculo de la similitud estructural conlleva la mayor parte de la carga computacional y por tanto el mayor tiempo de

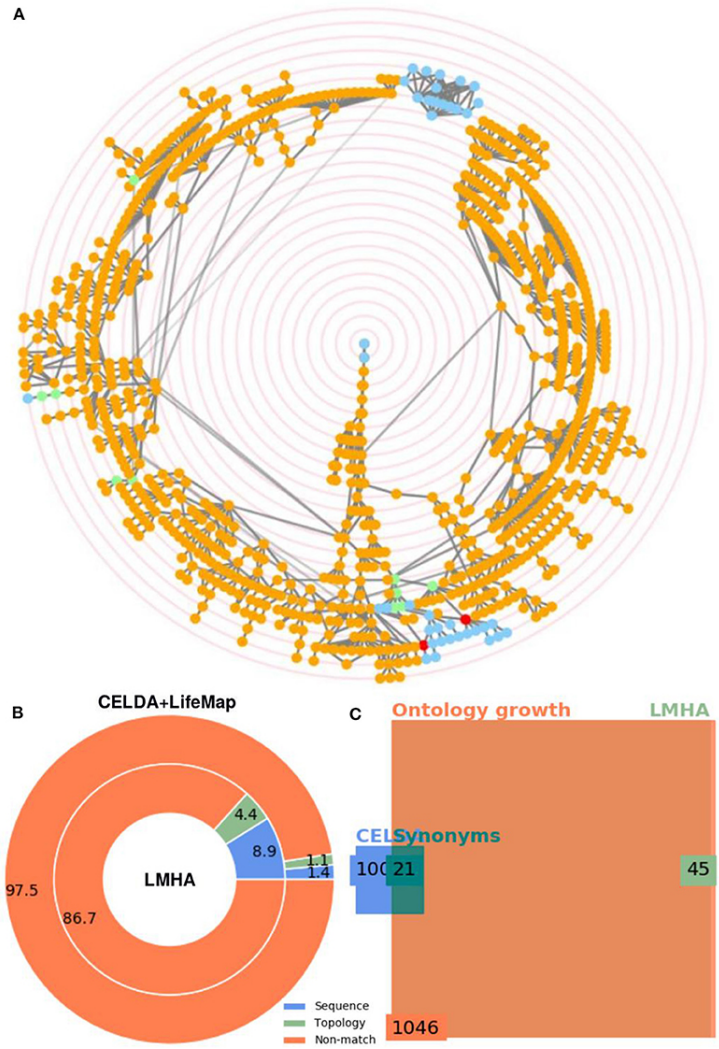


Figura 2.14: **Fusión de CELDA + LifeMap con la ontología LungMAP Human Anatomy (LMHA).** (A) Dígrafo circular acíclico (DAG) de la ontología fusionada. Los nodos naranjas y azules son la contribución no emparejados de CELDA+LifeMap y LMHA, respectivamente. Los nodos verdes y rojos son los nodos con emparejamiento por nombre y por estructura, respectivamente. (B) Donut plot de los porcentajes de clases agregadas mediante mapeo por nombre frente a las clases agregadas mediante mapeo de estructura a la ontología de CELDA + LifeMap (circulo externo) y LMHA (circulo interno). (C) Diagrama cuadrado de (B) Diagrama de anillos de los porcentajes de clases agregadas por mapeo de nombres frente a las clases agregadas por mapeo de estructura al CELDA + LifeMap combinado (círculo exterior) de LMHA (círculo interior). (C) Diagrama de cuadrados 'Euler-Venn' con el número de clases antes y después de la fusión. Los rectángulos azul y verde representan el número de clases de CELDA+LifeMap y LMHA antes de la fusión, respectivamente. El rectángulo verde oscuro enmarca la suma de clases emparejadas mediante nombre y estructural y el rectángulo naranja enmarca el número total de clases que obtiene la ontología fusionada de CELDA + LifeMap + LMHA. Los parámetros de alineamiento son: $W = 4\,\theta_{LN} = 0{,}7$ y $\Theta_N = 0{,}85$

cálculo del algoritmo por lo que se han implantado tres métodos para llevarlo a cabo:

- Similitud topológica vectorial. El cual incluye un método general para calcular las similitudes entre dos vectores de diferentes tamaños aplicando diferentes métricas
- Similitud topológica basada en restricción.
- Similitud topológica basada en el método de Blondel.

Diferentes fusiones de ontologías pueden obtener mejores resultados en función del uso de diferentes parámetros y métricas. Para el caso de la fusión de CELDA y LifeMap se observó que los métodos vectoriales producen resultados similares, con una ligera ventaja del método coseno. Todas las funcionalidades de FOntCell permiten la unificación de conocimiento disperso de un dominio en una única ontología. FOntCell produce los resultados en un formato típico de ontologías que puede ser reutilizado por el propio FOntCell de forma iterativa adaptando continuamente las ontologías a la luz de nuevos datos de ese dominio del conocimiento. Produce también un documento HTML con el que se puede observar los resultados de la fusión.

FOntCell es una herramienta desarrollada expresamente para la fusión de ontologías de desarrollo celular. El objetivo detrás de esta herramienta es la obtención de una ontología de tipo celular, el cual base sus relaciones en el proceso natural de diferenciación y desarrollo celular que sirva como punto de partida para la construcción del paisaje de Waddington a realizar en los siguientes capítulos.

FOntCell permite la recolección de información de diferentes ontologías de tipos celulares y contrastarlas entre ellas sin una supervisión estándar y nos garantiza una ontología resultante que contenga tanto los tipos celulares que hay en común como aquellos que no estén en común en ambas ontologías. Esta herramienta al ser creada con este objetivo en el horizonte no se obtienen los mismos resultados al fusionar otros tipos de ontologías que tengan otro tipo de jerarquía interna. Por lo que al intentar fusionar ontologías de dominios diferentes se obtienen resultados por debajo de la media respeto al resto de herramientas del OAEI. También, es cierto, que dentro los algoritmos menos específicos que son capaces de alinear ontologías, existen algunos que lo hacen mejor que FOntCell.

Capítulo 3

Etiquetado de muestras de bases de datos transcriptómicas mediante técnicas *big data*

3.1. Introducción

De cara a la construcción del paisaje de Waddington el primer paso ha sido la elaboración de una ontología de tipo celular basada en el desarrollo Cabau-Laporta et al. (2021) y que establezca el entramado topológico del paisaje de Waddington. El siguiente paso es la obtención de los datos transcriptómicos disponibles asociados a los tipos celulares de la ontología, para identificar los valores numéricos que darán forma al paisaje de Waddington.

Hemos abordado la construcción del paisaje de Waddington a partir únicamente de datos transcriptómicos, por la baja disponibilidad de datos epigenómicos que abarquen gran variedad de tipos celulares cuando se comenzó la presente tesis. Por ello, se definió el objetivo 4 como la construcción del paisaje transcriptómico de Waddington. Para realizar dicha construcción debemos buscar datos sobre transcriptómica asociada a tipos celulares, y adicionalmente, con información del estado transcriptómico de todos los genes de cada muestra. En la actualidad existen numerosas tecnologías para observar el estado transcriptómico de una célula (*single-cell analysis*), o de un conjunto presumible de un mismo tipo celular o de tejidos concretos (*bulk-cell analysis*).

Un escenario ideal para abordar nuestro estudio sería la producción de datos en masa de todos los tipos celulares implicados, lo que conllevaría establecer un protocolo donde se debería tomar muestras, cultivar, reproducir y analizar los diferentes tipos celulares resultantes de la fusión de

las ontologías de CELDA y LifeMap, elevando los costes y tiempos de procedimiento del presente estudio.

Para ahorrar costes y reutilizar la gran cantidad de datos transcriptómicos existentes hemos optado por la obtención de los mismos procedentes de repositorios *online*, lo cual otorga ciertas ventajas y plantea otros retos. En primer lugar, ofrece la ventaja de ahorrar los costes asociados a la experimentación, obtención de muestras, etc, y permite realizar el proyecto sin la necesidad de realizar experimentos *wet lab*.

Por otra parte, el uso de datos alojados en repositorios *online* brinda la oportunidad de reutilizar y reanalizar una serie de datos ya publicados por otros grupos en el contexto de otros experimentos, lo cual, permite utilizar el modelo molecular que deseemos. Esto no solo supone un ahorro general, sino que también el reciclado de estos datos permite un control de calidad adicional y masivo sobre los mismos.

El uso de datos de diferentes experimentos plantea también dos retos principales: por un lado el efecto de la influencia de las distintas características de los lotes de los datos (*batch effect*). Por otro lado, la gran disponibilidad de datos hace prácticamente imposible el etiquetado manual de los mismos y obliga al desarrollo de algoritmos *Big-data* para el etiquetado masivo automático y la búsqueda concreta de tipos celulares.

3.1.1. Naturaleza de los datos transcriptómicos utilizados

Dada la gran disponibilidad de los datos alojados en la base de datos alojados en GEO a la hora de desarrollar los algoritmos necesarios para construir el paisaje de Waddington se optó por el uso de datos alojados en esta base de datos y más concretamente para *Affymetrix* GPL570 de Humano que es la que más datos de distintos tipos celulares humanos contenía a fecha de elaboración del presnete trabajo.

GEO nace con la intención de proporcionar un repositorio público de datos de expresión genética (Edgar et al., 2002). En GEO la información está organizada en una de las 3 categorías siguientes:

- *Platform* (GPL): La plataforma define el tipo de moléculas que se detectan. Esto proporciona información de la técnica utilizada, el tipo de dato obtenido y del organismo sobre el que se utiliza (Edgar et al., 2002). Aquí se ha utilizado la plataforma GPL570, cuyo identificador se refiere a *Affymetrix Human Genome U133 Plus 2.0 Array* la cual es el conjunto más completo de los *GeneChips* de *arrays* de oligonucleótidos que ofrece *Affymetrix* para la evaluación de la transcripción del genoma completo (Edgar et al., 2002). El prefijo 'GPL' es el encargado de definir las diferentes plataformas de GEO, seguido de una serie de números indicativos,

como por ejemplo la plataforma GPL570 (GPL+570) (Edgar et al., 2002).

- *Series* (GSE): Las series se refieren a los diferentes experimentos realizados en el seno de una plataforma. Los experimentos pueden tener asociados más de una plataforma diferente, dependiendo del trabajo. Los experimentos tienen asociados principalmente una serie de muestras correspondientes a ese experimento (Edgar et al., 2002). El prefijo GSE es el encargado de definir los diferentes experimentos o *series*, como por ejemplo el GSE1145 (GSE+1145) (Edgar et al., 2002).

- *Samples* (GSM): Las muestras se refiere al conjunto de datos obtenidos de cada test individual dentro del análisis con una única plataforma (Edgar et al., 2002). Por tanto, las muestras solo tendrán definidas una única plataforma (Edgar et al., 2002), aunque podrán estar adscritas a más de un experimento si han sido utilizadas en más de uno. En nuestro caso, cada muestra se refiere a una matriz con la expresión transcriptómica de todo el genoma de células del mismo tipo analizadas a la vez. El prefijo referido a las muestras es GSM y como en los casos anteriores viene definido con un conjunto de números identificativos de cada muestra, por ejemplo, la muestra GSM18422 (GSM + 18422) (Edgar et al., 2002).

Cada una de estas entidades tiene su propia página en la web de GEO desde la cual se puede observar diferentes metadatos con información relevante al tipo de experimento, en el caso de los GSEs, y al tipo de muestra, en el caso de los GSMs.

Trabajar con los datos de una plataforma conlleva la incorporación a un mismo análisis datos provenientes de dos o más GSEs lo cual implica el tratamiento del *batch effect*. El *batch effect* es el fenómeno que sucede al analizar diferentes muestras de diferentes trabajos o lotes de experimentos, en los cuales las muestras tienden a agruparse principalmente por experimentos o por laboratorios, dificultando así la posibilidad de comparación entre muestras de diferentes experimentos(Irizarry et al., 2003). Estas diferencias de expresión entre muestras son debidas a variaciones introducidas durante la preparación, creación de los *arrays* y la propia hibridación (Irizarry et al., 2003).

Dos muestras de un mismo tipo celular, pero de diferente GSE, pueden ser más similares entre ellas que con muestras de diferentes tipos celulares, pero del mismo GSE, que entre ellas. En este contexto, es necesario aplicar algún algoritmo que mitigue en la medida de lo posible este efecto, permiten así evitar resultados que segreguen dependiendo del experimento y no del tipo celular.

El procesamiento habitual para los datos de *microarrays* de *Affymetrix* consiste en la normalización conjunta de los datos mediante el algoritmo *Robust Multichip Array* (RMA) (Irizarry et al., 2003)). Esta normalización (junto con la alternativa GC-RMA, que hace una corrección de conte-

nido GC de los *probes*) suele estar implícito en todos los protocolos que trabajen con datos de *Affymetrix*, incluso entre aquellos que trabajen con datos de un mismo y único experimento.

Aunque la normalización puede contribuir al a reducción del *batch effect* no siempre consigue eliminarlo totalmente, sobre todo entre muestras de diferentes experimentos. Con esto en mente es necesario aplicar técnicas de reducción del *batch effect* como *ComBat* (Johnson, Li, 2007). *ComBat* es un algoritmo basado en métodos de Bayes empíricos (Johnson, Li, 2007). Este algoritmo recibe como entrada los datos y la especificación por parte del usuario de los qué lotes se han utilizado para así eliminar su efecto. Esta supervisión por parte del usuario permite al algoritmo conocer como escalar el conjunto de datos de modo que se minimice la asociación de los datos en función de estos lotes. *ComBat* devuelve una matriz de datos transformada de modo que el resto de información se mantenga en la medida de lo posible y la información puramente asociada a cada lote se diluya.

3.1.2. Selenium para el etiquetado masivo de tipos celulares alojados en GEO

GEO contiene datos transcriptómicos de muchos y diferentes tipos celulares en el contexto de diferentes experimentos. Pero uno de los principales retos a abordar es su correcto etiquetado. Para ello se debe atender a los metadatos asociados con cada muestra.

La plataforma GPL570 contenía (a fecha 30/12/2022) 167484 GSMs repartidos en 5562 GSEs. Los metadatos más importantes en cada muestra están repartidos en torno a 5 campos de entrada (**Tabla 3.1**). Por lo que la cantidad de metadatos a analizar asciende a un total de 836355 Teniendo en cuenta que para etiquetarlos hay que enfrentar la totalidad de esos metadatos a la totalidad de las posibles etiquetas (los distintos tipos celulares) implica la construcción de una matriz de 1177587840 celdas diferentes.

Realizar este etiquetado manualmente resultaría una tarea inabarcable, la enorme cantidad de datos alojados en GEO, para una única plataforma podría encajar dentro de la debatida definición de *Big data*;(Cappa et al., 2020).

Existen técnicas que pueden ayudar al análisis de *Big data* alojado en una página web, uno de ellas es la técnica del *web scraping*;(Zhao, 2017). Esta técnica se basa en la obtención y construcción de bases de datos a partir de la información disponible en una web principalmente de forma automática mediante el uso de un *web crawler/web driver*, es decir, un *bot*. Pero también puede realizarse de forma manual a través de un usuario el cual interprete el código *html* de la página en cuestión y extraiga de ahí la información de interés (mediante el uso de *web-parsers*, los cuales ayudan a la interpretación del código *html*) (Zhao, 2017). En este caso, debido a la cantidad de información que queremos extraer nos decantaremos por el uso del *web driver*.

El *web scraping* se divide en dos etapas, en la primera se accede a una determinada página web y se extrae la información relevante o incluso toda la información de la web (extracción de recursos web) y durante la siguiente etapa se ordena y se construye un *dataset* con la información descargada o parte de la misma (extracción de información) (Zhao, 2017).

Existen numerosos programas con diferentes opciones y funcionalidades para la construcción y automatización del *web scraping*. Entre ellos: BeautifulSoup, Selenium e Urllib2, entre los más comunes (Zhao, 2017).

Selenium tiene una implementación que permite realizar simultáneamente las dos tareas definidas del *web scraping*. Además, lo realiza mediante la creación de un *web driver* automatizado que permite automatizar la navegación web (Zhao, 2017);(Nyamathulla et al., 2021), adicionalmente permite la implementación en Python.

El *web driver* de Selenium genera un *bot* de un determinado navegador que puede programase con una serie de órdenes que se irá ejecutando secuencialmente (Nyamathulla et al., 2021). Dada la arquitectura de la web de GEO y de NCBI, y de la cantidad de entradas a analizar esto lo convierte en una herramienta útil para algunos de nuestros algoritmos de etiquetado de muestras.

3.2. Materiales y métodos

3.2.1. Etiquetado y obtención de las muestras

Se parte de la lista de tipos celulares obtenida mediante la fusión de las Ontologías de CELDA y LifeMap utilizando la herramienta para alineamiento y fusión de ontologías: FOntCell. Esta lista contiene los tipos celulares consenso de ambas ontologías, manteniendo todas las relaciones que tienen entre ellas. CELDA y LifeMap poseen información principalmente de humano y ratón, en la fusión realizada en los pasos anteriores ya se preseleccionaron las clases de ambas ontologías relativas a datos humanos, por lo que se dispone de una base de datos de tipos celulares humanos.

El objetivo es extraer la información necesaria a partir de los metadatos que estén disponibles en la web de GEO en lo referente a cada GSM para poder así identificar cada GSM con un determinado tipo celular de nuestra ontología, siempre y cuando sea posible. Estos metadatos se generan mediante las entradas que introduce el usuario cuando sube los datos a GEO.

De estos metadatos se deduce, en función de cómo esté redactada y expuesta la información de la muestra, el tipo celular. Pero dada la inmensa cantidad de metadatos que hay en una sola plataforma se decidió utilizar algoritmos *Big-data* para el *web scrapping* y para ello se plantearon

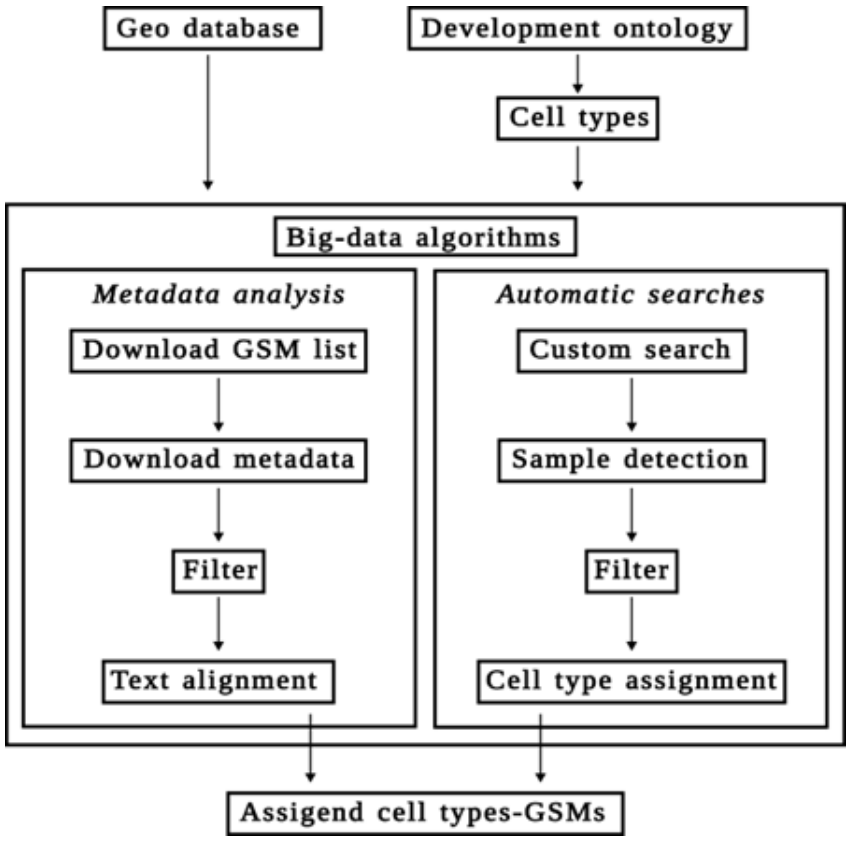


Figura 3.1: **Diagrama de flujo de los algoritmos *Big-data* creados para la asignación de tipos celulares a los GSMs de la plataforma GPL570 de Affymetrix.**

dos algoritmos funcionando en paralelo y desde dos puntos diferentes para poder etiquetar el máximo número posible de muestras de GEO.

3.2.2. Algoritmo de análisis de metadatos

Asumimos que si un determinado tipo celular es nombrado con mayor frecuencia de los metadatos de una determinada muestra y/o el tipo celular aparece nombrado en unas determinadas entradas de los metadatos la muestra corresponderá a ese determinado tipo celular. El diagrama de flujo asociado a este algoritmo se presenta en la **Fig. 3.1**.

La propia plataforma de GEO facilita la posibilidad de descargar un documento con todos los nombres de las muestras listado para una determinada plataforma, en nuestro caso GPL570. Dada la propia arquitectura de la web de GEO, conociendo el nombre de un GSM se puede generar el *link* que llevará a la página específica del GSM en cuestión. Atendiendo a estas dos cuestiones se accede a cada una de las páginas de cada GSM. Y de cada una de las páginas de cada GSM se descargan los metadatos contenidos en los campos considerados relevantes en la **Tabla 3.1**.

Entrada de Metadato	Disponibilidad del metadato	Ponderación	Aplicación de filtro
Title	Usualmente presente	1	Si
Source name	Usualmente presente	1	Si
Description	Usualmente presente	0.5	No
Label	Raramente presente	1	Si
Characteristics	Raramente presente	0.5	No
Summary	Raramente presente	0.5	No

Tabla. 3.1: **Tabla de tratamiento de metadatos de GEO para la plataforma GPL570 de Affymetrix**. Tabla donde se encuentra las principales entradas de los metadatos de cada muestra que contienen la información más relevante respecto del tipo celular. En la primera columna se listan los nombres de entradas seleccionadas. En la segunda la disponibilidad de esos metadatos a lo largo de las diferentes muestras. En la tercera se añade la ponderación de las diferentes entradas, para reflejar cuales tienen más peso a la hora de calcular la puntuación de los tipos celulares que se encuentren a lo largo de los metadatos de un determinado GSM. En la cuarta columna se indica aquellas entradas de metadatos las cuales son filtradas si en ellas aparece una de las siguientes palabras: {cancer, cancerous, immortalized, immortal, derived, cell-derived, derived-cell }.

El siguiente paso consiste en generar un algoritmo que interprete los textos de los metadatos descargados. Durante el desarrollo de FOntCell (Cabau-Laporta et al., 2021) creamos algoritmos y métricas de similitud entre dos cadenas de texto. Con estas métricas podemos buscar en el conjunto de textos, palabras similares a grupos de palabras clave (como podrían ser los tipos celulares) y a partir de su presencia o no, tomar ciertas decisiones.

Nuestro algoritmo averigua los tipos celulares que son nombrados en los metadatos asociados a un determinado GSM, en función de la ponderación (de qué metadato proviene) expuesta en la (**Tabla 3.1**) y la repetición se elabora una puntuación. El tipo celular que tenga la mayor puntuación con ese determinado GSM será el que finalmente se asigne. La ponderación que se muestra en la (**Tabla 3.1**) representa la importancia de un término encontrado en una entrada u otra, por ejemplo, usualmente las entradas de *Description* o *Characteristics* suelen describir términos más generales del experimento y de las circunstancias de la muestra, sin embargo, el *Tittle* o el *Source name*, si contienen un nombre identificable de un tipo celular se puede afirmar que la muestra versará de ese tipo celular. Por ello la ponderación otorga la mitad del peso encontrar el nombre de un tipo celular en los campos de *Description* y *Characteristics*.

Para establecer una comparativa entre las cadenas de texto de los metadatos y de la ontología de desarrollo celular se utiliza la misma lógica desarrollada en FOntCell para el alineamiento de ontologías por nombre (**Fig. 3.2**). Se realiza una división del texto de los metadatos de acuerdo a una ventana convolutiva de palabras de tamaño igual a la cadena de texto del tipo celular con el que se está alineando. Se generan todas las posibles combinaciones de texto que tengan ese tamaño y se genera una matriz de similitudes, si existen similitudes por encima del valor 0.85 se asignará ese tipo celular como candidato y en función de la ponderación se añadirá un 0.5 o un 1.

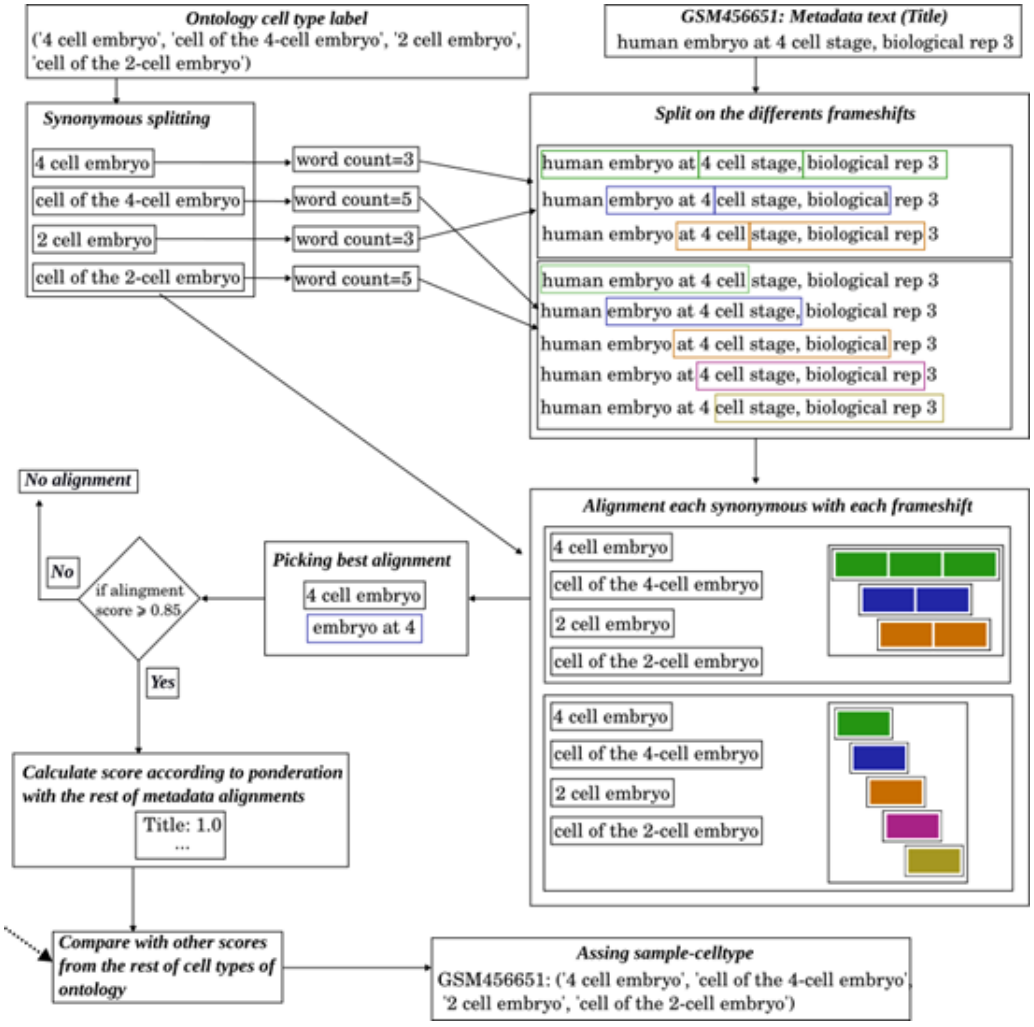


Figura 3.2: **Diagrama de flujo del proceso de alineamiento de sinónimos entre la base de datos y los metadatos extraídos de la web.** Se muestra como el tipo celular de la ontología es descompuesto en sus sinónimos, después se calculo el número de palabras de cada sinónimo y con ellos se crean los marcos de lectura, en este caso: un marco de lectura que agrupa las palabras del metadato del GSM en conjuntos de 3 palabras y otro marco que las agrupa en conjuntos de 5 palabras. A continuación los marcos de lectura son alineados con los sinónimos y mediante el cálculo la distancia de de Levenshtein Levenshtein (1966) se selecciona el mejor emparejamiento. Se realiza el mismo proceso para el resto de metadatos del GSM y se almacenan en 'valores absolutos' (de acuerdo con la **Tabla 3.1** aquellos que superan el umbral de similitud asignado. Finalmente se establece una puntuación global con el resto de puntuaciones de otras clases de la ontología para seleccionar como asignación aquella que tenga una mayor puntuación.

A continuación, la suma de puntuaciones de todas las entradas será la puntuación relativa de ese GSM con el tipo celular concreto. Se asignará entonces, aquel tipo celular con mayor puntuación siempre y cuando tenga una puntuación mínima de 1, para evitar falsos positivos de tipos celulares nombrados exclusivamente en *Characteristics* y *Description*.

Un paso previo a la elaboración de la puntuación y el etiquetado, consiste en filtrar aquellas muestras que hagan referencia a condiciones de las muestras que no deseables de ser introducidas en nuestro sistema, es decir, aquellas que contienen palabras como las mostradas en (**Tabla 3.1**) en las posiciones ponderadas con 1, es decir, en *Tittle*, *Source name* o *Label*.

Recapitulando, el Algoritmo de análisis de metadatos recorre cada uno de los GSMs, descargando los metadatos, para a continuación aplicar el filtro de las palabras clave mostradas en la **Tabla 3.1**. Después, el algoritmo realiza un alineamiento entre los metadatos descargados y la base de datos de tipos celulares generando las asignaciones que considere pertinentes de acuerdo a las normas y parámetros establecidos. Finalmente, se obtiene una lista de tipos celulares presentes en nuestra ontología con los GSMs asignados y los GSEs a los que pertenecen.

3.2.3. Algoritmo de búsquedas automáticas

Una forma de encontrar tipos celulares concretos alojados en GEO consiste en realizar búsquedas personalizadas en el buscador de NCBI. En nuestro caso, para buscar muestras de la plataforma de *Affymetrix* GPL570 se debe introducir la siguiente estructura: GPL570[Accession]+'nombre del tipo celular'.

El motor de búsqueda de NCBI muestra tras una búsqueda personalizada con la estructura mencionada, los GSEs y GSMs que tengan algún tipo de vinculación con el tipo celular que se haya introducido. Los resultados se muestran en la página principal y por defecto muestra 20 resultados por página, hasta completar el total de GSEs (listados inicialmente) y GSMs (listados a continuación de los GSEs) vinculados a ese tipo celular. Los vínculos están definidos en función de los metadatos asociados no solo al GSM si no también al GSE. Parte de estos metadatos son mostrados en los resultados de la búsqueda en NCBI. En la (**Fig. 3.3**) se muestra un ejemplo de los resultados obtenidos de una búsqueda, resaltando las entradas importantes para nuestro algoritmo.

El Algoritmo de búsquedas automáticas pretende automatizar el proceso de búsquedas personalizadas en NCBI y seleccionar aquellas muestras que se consideren más relevantes, el funcionamiento se ilustra en la (**Fig. 3.1**).

La arquitectura de la web de NCBI genera un *link* específico vinculado a cada búsqueda que facilita

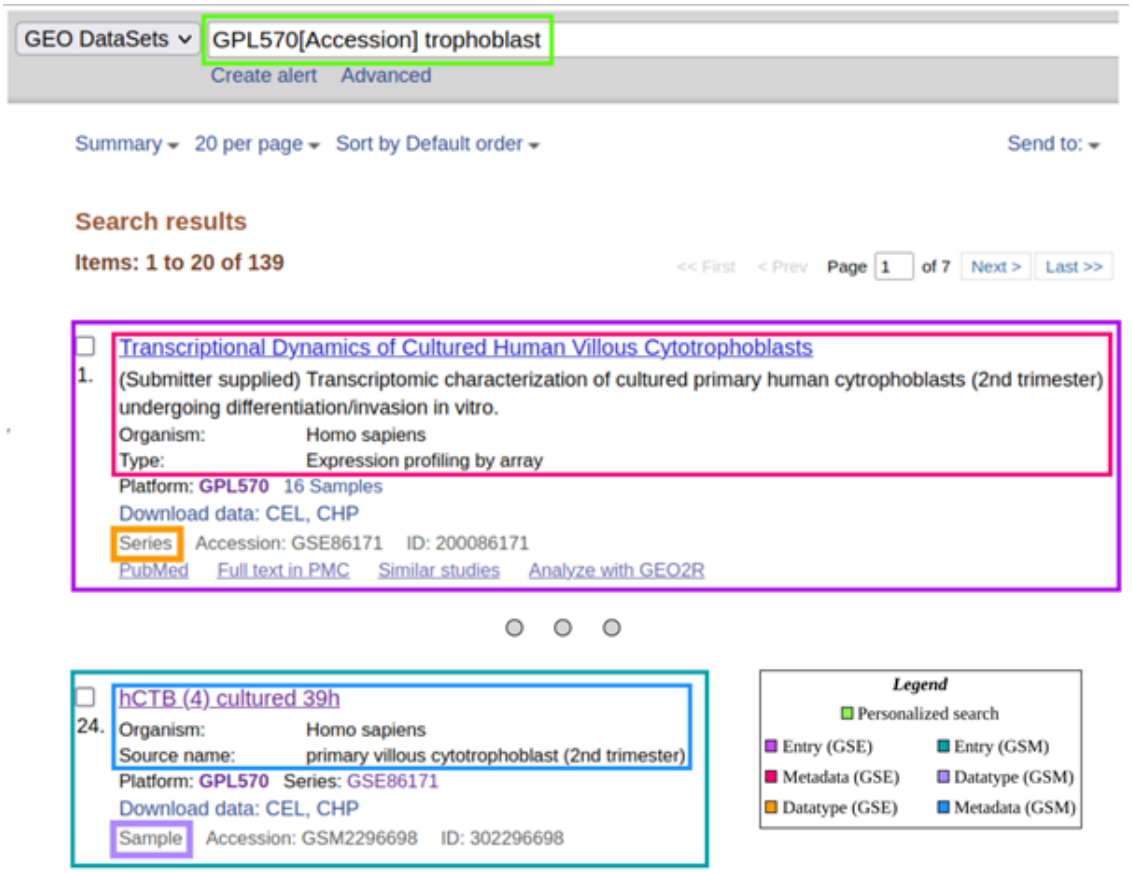


Figura 3.3: **Ejemplo de una búsqueda personalizada en GEO.** En este ejemplo se ha lanzado la búsqueda personalizada para la plataforma GPL570 y el tipo celular 'trophoblast'. En los resultados de la respuesta del servicio *web* se muestra el primer resultado, correspondiente a un GSE y adjuntado el resultado número 24 el primer resultado correspondiente a un GSM. En ambos resultados se observan la información que utiliza el algoritmo, correspondiente a los metadatos y el tipo de resultado: '*Series*' correspondiente a un GSE o '*Sample*' correspondiente a un GSM.

la tarea de realizar búsquedas automáticas ya que posibilita generar un *link* específico a cada una de las búsquedas, es decir, un *link* para cada uno de los tipos celulares.

A continuación, utilizando la librería de Selenium Nyamathulla et al. (2021) en python se crea un *bot* para el buscador de Firefox Web Browser. El *bot* tiene las instrucciones de acceder a cada uno de los *links* generados a partir de los tipos celulares. El acceso se realiza de forma secuencial, es decir, solo se accede a la búsqueda de un tipo celular al mismo tiempo con el objetivo de no saturar la entrada web o de no ser considerado como un ataque informático de denegación de servicio (DoS) y, por tanto, el administrador de la *web* NCBI corte la conexión.

Después de lanzar una búsqueda, se recopila la información relevante a cada una de las entradas de las búsquedas y se navega por cada una de las pestañas hasta finalizar todas las páginas generadas por dicha búsqueda.

El *bot* recopila la información relevante para categorizar los diferentes resultados de la búsqueda, resaltados en la (**Fig. 3.3**). De la información de los resultados de las búsquedas se pueden seleccionar aquellos resultados referentes a muestras. Una vez localizado las muestras listadas, se realiza un pequeño filtro en función de los metadatos que se muestran en los resultados de la búsqueda.

Tal y como se muestra en la (**Fig. 3.3**), en el caso de las muestras aparecen los metadatos relativos al 'Source Name' y 'Title'. En esos metadatos se aplican los filtros con las palabras claves mostrados en la (**Tabla 3.1**), para eliminar posibles muestras de datos no deseadas de introducir en nuestro *pipeline* de construcción del paisaje de Waddington. Una vez aplicado el filtro, se asignan las muestras no filtradas al tipo celular buscado.

En resumen, el Algoritmo de búsquedas automáticas utiliza las 1408 muestras de la Ontología resultante de FOntCell para generar enlaces de búsquedas del buscador de NCBI, utilizando la estructura: GPL570[Accession]+'nombre del tipo celular', después se recorren las distintas páginas mostradas con los resultados de la búsqueda. En cada página se observa cada uno de los resultados para primero identificar si es una muestra, después observar en los metadatos la ausencia de las palabras clave mostradas en la **Tabla 3.1** y finalmente asignarlo al tipo celular de la búsqueda que ha generado el enlace.

3.2.4. Métrica de cobertura de la existencia datos transcriptómicos de tipos celulares

Una vez etiquetadas las muestras es importante estudiar el grado de cobertura de un árbol de desarrollo de tipos celulares conectados en función de la ascendencia-descendencia generado

por FOntCell. Para medir dicha cobertura de regiones celulares con disposición de datos transcriptómicos creamos una métrica a la que llamamos Cobertura. La Cobertura es una métrica de distribución de los tipos celulares a lo largo de la ontología modelo. En primer lugar, reutilizando parte del software desarrollado durante la creación de FOntCell, a lo que procesamiento de grafos se refiere, generando una serie de subgrafos de distinto tamaño de ventana w. Así, se genera un subgrafo por cada tipo celular de la ontología como lo hace FOntCell. A continuación, se calcula el porcentaje de subgrafos de los cuales existen datos transcriptómicos en al menos un tipo celular del conjunto de tipos celulares que conforman cada subgrafo.

Esta métrica mide el porcentaje de creodos de tamaño w de los cuales se dispone de datos a lo largo de toda la distribución de tipos celulares. Por ejemplo, una Cobertura para $w = 3$ implicaría que tomando todas las relaciones entre los tipos celulares del estilo: 'abuela-madre-hija' si el valor resultante fuese del 50 % se puede afirmar que para todas las estructuras de este tipo que encontramos a lo largo de la ontología se dispone de datos para al menos un elemento en el 50 % de los casos.

Esta métrica es sensible al tamaño de w, al incrementar dicho tamaño de ventana se añaden ascendentes/descendentes al conjunto total, otorgando cada vez estructuras más grandes e incrementando por tanto la posibilidad de encontrar información. Por otra parte, un tamaño de ventana $w = 1$ implica una estructura de tipos celulares únicos, lo que equivale al porcentaje de tipos celulares con datos transcriptómicos respecto del total de tipos celulares de la ontología.

3.2.5. Normalización de los datos y reducción de *batch effect*

La normalización realiza diferentes transformaciones matemáticas con el objetivo de mantener las relaciones numéricas entre las muestras, pero eliminar las diferencias de escala que entre ellas. Este proceso permite ajustar los datos a una misma escala, eliminar parte de los sesgos, y reducir el *batch effect*;(Johnson, Li, 2007).

Utilizaremos el algoritmo de normalización RMA del paquete Bioconductor de R. Este algoritmo consiste en una secuencia de operaciones (Wagner, 2016) partiendo de los datos crudos (ficheros .CEL) de *Affymetrix* de los que se obtiene una matriz de intensidades para cada conjunto de *probes* (representando un gen) de cada muestra. El conjunto de operaciones del algoritmo de normalización RMA es:

1. Descartar las *mismatch probes* (MM *probes*), es decir, aquellas *probes* en las que, habiendo una hibridación, tienen algún nucleótido no coincidente. Eliminar estas *probes* genera medidas de la expresión más precisas (Irizarry et al., 2003).

2. Eliminación del ruido de fondo. El ruido de fondo se elimina a partir de la distribución de intensidades observadas, generando una variable aleatoria que representa la suma del ruido de fondo con una distribución normal truncada (Wagner, 2016).

3. Normalización de cuantiles a la matriz de intensidad (Irizarry et al., 2003). Aunque no tiene mucho efecto a la hora normalizar muestras similares, puede tener un mayor efecto si se utilizan datos más heterogéneos (Wagner, 2016).

4. Transformación logarítmica de todos los valores de la matriz de intensidad. Este paso no es robusto a los valores inferiores a 1.

5. Agrupación los *probes* con el gen al que pertenecen generando una submatriz de intensidades por gen. En estas submatrices se aplica la media polaca (Tukey, 1977) que genera un único valor de expresión para cada gen (o por varias isoformas del gen) a partir de varias *probes*; (Wagner, 2016)

Para la construcción final del paisaje de Waddington se utiliza todo el conjunto de muestras etiquetadas, pero para las pruebas de concepto de reducción del *batch effect* y de los algoritmos de construcción del paisaje de Waddington se ha seleccionado un sub-conjunto conocido de muestras de hematopoyesis.

Este conjunto consiste en 3 GSEs de datos de hematopoiesis (GSE42519, GSE49910, GSE123991). Se dispone de 95 muestras cubriendo 12 tipos celulares hematopoyéticos, de los cuales 5 están presentes en más de un GSE. Utilizar datos combinados de 3 experimentos diferentes permite comprobar el grado de funcionamiento de los algoritmos para la Reducción del *batch effect*.

La eficacia de la eliminación del *batch effect* se evidencia al conseguir que muestras de un mismo tipo celular se agrupen independientemente del experimente (GSE) de origen, es decir, se presenten más agrupadas entre sí que con otras muestras del mismo GSE y por tanto las muestras de un mismo GSE correspondientes a distintos tipos celulares deben tender a alejarse tras aplicar técnicas de reducción del *batch effect*. Es decir, en la prueba de concepto analizamos si *ComBat* consigue eliminar el *batch effect* que que contiene los datos de un grupo de experimentos (GSEs en este caso) y que se mantengan, e incluso aparezcan, nuevas agrupaciones fruto de las similitudes que tengan las muestras de un mismo tipo celular independientemente del GSE.

Con la idea de obtener una métrica de la agrupación se calcularon las medias de las distancias en el análisis de componentes principales (PCA) de una muestra de un determinado tipo celular con el resto de muestras del mismo tipo celular, independientemente del GSE. También se calculó la distancia media de una muestra aleatoria de cada GSE con el resto de muestras de ese mismo GSE.

Las distancias medias se calcularon obteniendo las distancias de todos los puntos de un mismo grupo (GSE o tipo celular) a un punto de ese grupo y a continuación calculando la media aritmética de todas las distancias de un mismo grupo.

$$Distancia = \sqrt{(X_0^2 - X_1^2) + (Y_0^2 - Y_1^2)} \quad \text{(Ec. 3.2.1)}$$

Donde X_0 y Y_0 son coordenadas cartesianas del punto 0 y X_1 y Y_1 las coordenadas cartesianas del punto 1. Por lo que aplicando el teorema de Pitágoras conocemos la distancia entre ambos. De esta forma calculamos la distancia de cada punto i al punto de referencia y obtener la media de las distancias aplicando la ecuación:

$$Media\ distancias = \frac{1}{Total\ distancias} \sum_{i=1}^{Total\ distancias} distancia_i \quad \text{(Ec. 3.2.2)}$$

Donde se calcula el sumatorio de todas las distancias i y se divide para $Totaldistancias$ que es un número natural igual al número de distancias que disponemos (que será igual al número de puntos eliminado al punto de referencia).

Aplicando estas ecuaciones obtenemos una distancia media ($Mediadistancias$ en la 3.2.2 de las muestras de cada tipo celular, y una distancia media de las muestras de cada uno de los GSEs. Estas métricas se realizan antes de utilizar *ComBat* (únicamente habiendo aplicado la normalización RMA) y después de utilizar *ComBat* como técnica de reducción del *batch effect*. De esta forma se puede medir como afectan las técnicas de reducción del *batch effect* a la agrupación de los datos respecto al tipo celular y como afecta esto a los lotes originales.

En aquellos casos que las técnicas de reducción del *batch effect* han disminuido la distancia media de los puntos de un tipo celular se puede afirmar que las técnicas de reducción del *batch effect* han mejorado la agrupación por tipo celular. Después se estima el porcentaje de tipos celulares cuya agrupación mejora tras el uso de técnicas de reducción del *batch effect*. Adicionalmente se observa la variación de la distancia media de los lotes originales tras utilizar técnicas de reducción del *batch effect*, para este caso se deberían mantener a distancias similares o incluso alejarse.

Es de especial interés observar cómo se comportan los tipos celulares que tienen presencia en varios de los GSEs que utilizamos en la prueba de concepto ya que es en estos casos donde las técnicas de reducción del *batch effect* más puede incidir al tener muestras en los diferentes lotes.

3.3. Resultados

3.3.1. La combinación de los dos algoritmos de etiquetado de muestras etiqueta un 8 % de las muestras disponibles en GEO para la plataforma GPL570

De la combinación en la aplicación de los Algoritmos de Búsquedas Automáticas y del Algoritmo de análisis de metadatos, se ha etiquetado un total de 20770 muestras de las 167484 muestras alojadas en GEO para esa plataforma y accesibles desde el buscador de NCBI. Esto se traduce en un etiquetado en torno al 8 % de las muestras alojadas en GEO para la plataforma GPL570.

Del total de 1408 tipos celulares diferentes categorizados en la ontología resultante de FOntCell obtenemos un total de 131 tipos celulares, que tras la normalización termina siendo un total de 121 tipos celulares diferentes lo que corresponde a un 8,5 % (**Fig. 3.5**) respecto al total de tipos celulares categorizados en la ontología resultante de FOntCell.

En esta ontología, atendiendo a los distintos linajes celulares: cigoto, mesodermo, ectodermo y endodermo, diferenciamos 517 tipos celulares distintos en el mesodermo, 120 tipos celulares distintos para el endodermo, 723 tipos celulares diferentes identificados para el ectodermo y 48 tipos celulares diferentes identificados para el linaje cigótico.

Atendiendo a los linajes de los tipos celulares (**Fig. 3.5**) observamos que de la totalidad de los tipos celulares del linaje cigótico un 20.8 % tiene datos transcriptómicos asociados en el conjunto de muestras etiquetados, siendo este linaje, en proporción, el que mayor porcentaje de tipos celulares con datos tiene con un total de 10 tipos celulares identificados. El endodermo y el mesodermo son los siguientes linajes con un mayor porcentaje de tipos celulares cubiertos respecto al total, cubriendo un total de un 10 % y un 9.5 % respectivamente que corresponden al total de 12 tipos celulares identificados para el endodermo y 67 identificados para el mesodermo. Finalmente, para el linaje del ectodermo se obtuvieron un 5.6 % de datos de los tipos celulares que corresponde a un total de 29 tipos celulares identificados para este linaje.

El dígrafo mostrado en **Fig. 3.4** proporciona una idea aproximada a la distribución de los tipos celulares a lo largo del total de tipos celulares de la ontología resultante de FOntCell. En la **Fig. 3.6** se muestra la distribución representada en número de nodos (totales a la izquierda y con datos a la derecha) de una sección circular equivalente al dígrafo.

Las representaciones de la distribución se complementan con la métrica de Cobertura la cual proporciona una visión del grado de distribución de los tipos celulares a lo largo del dígrafo. Las métricas del Cobertura se establecieron con subgrafos de tamaño de ventana w de 3, 4 y 5. Obteniéndose los resultados de Cobertura:

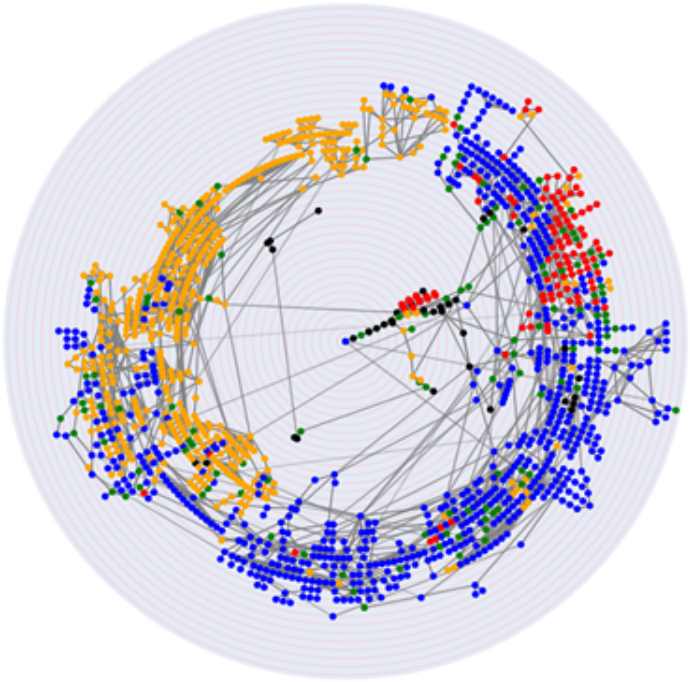


Figura 3.4: **Dígrafo del desarrollo celular**. En verde se muestran los tipos celulares de los cuales se dispone de datos. El negro, rojo, naranja y azul corresponde a los tipos celulares de uno de los diferentes linajes, cigoto, endodermo, ectodermo y mesodermo respectivamente.

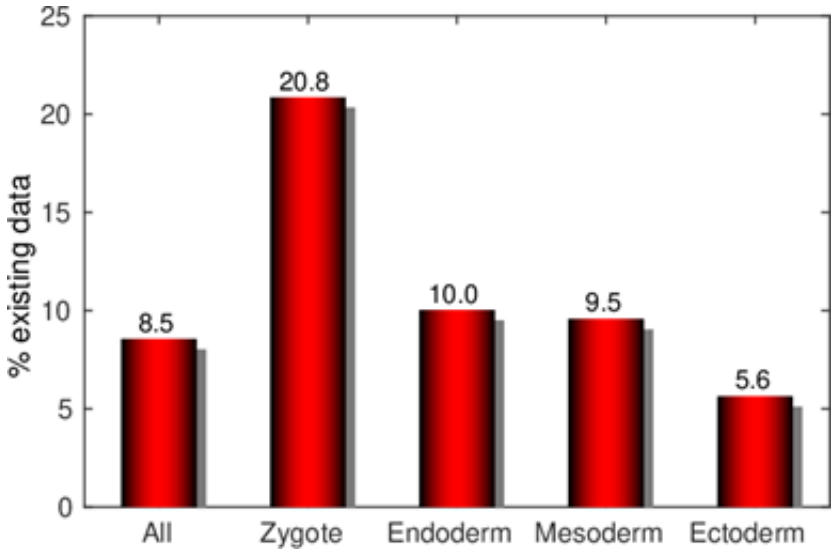
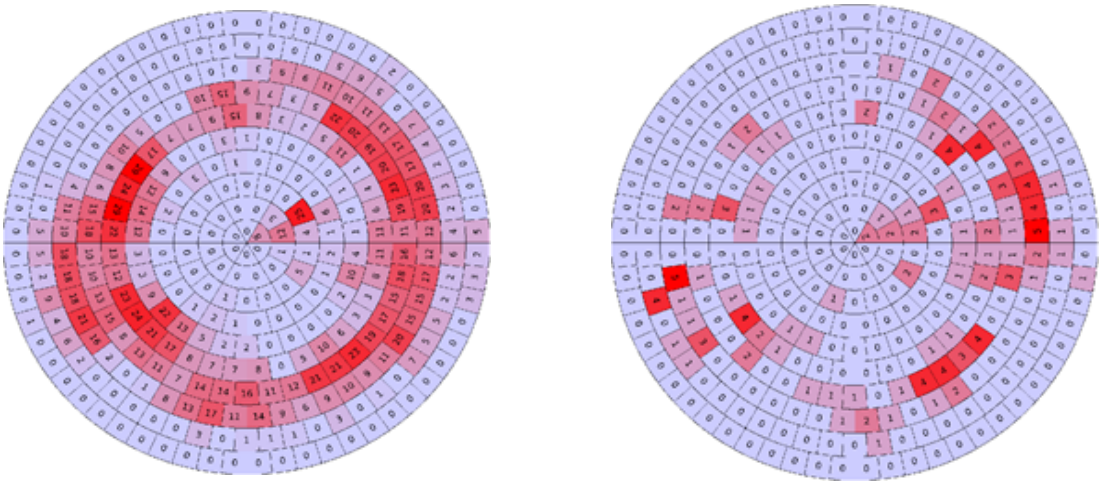


Figura 3.5: **Diagrama de Barras de los porcentajes de tipos celulares.** Generado con datos respecto al total y a cada uno de los diferentes linajes celulares.



(a) (b)

Figura 3.6: **Mapas de color circular de las diferentes áreas de los grafos circulares.** En cada una de estas áreas se contabiliza el número de tipos celulares. En el panel izquierdo se muestran las áreas contabilizando la totalidad de los tipos celulares en cada una de las regiones, en el panel derecho se contabilizan los tipos celulares para los cuales se dispone de datos transcriptómicos. La intensidad del color rojo indica el número de tipos celulares hallados para ese sector.

- Cobertura con w = 3: 41 %
- Cobertura con w = 4: 53 %
- Cobertura con w = 5: 64 %

A la luz de estos resultados se puede afirmar que se ha conseguido etiquetar un conjunto de tipos celulares distribuido a lo largo del dígrafo y que por tanto es posible interpolar un paisaje de Waddington representativo de la totalidad del desarrollo celular. Aunque el uso de un 8 % respecto a la totalidad de los datos públicos de GEO para la plataforma GPL570 pueda parecer pequeño, tenemos en cuenta que la mayoría de muestras alojadas en esta base de datos corresponde a tejidos y tipos celulares que son descartados por nuestros algoritmos. Estas muestras han sido descartadas puesto que en sus metadatos incluían indicios suficientes para que nuestros algoritmos identificaran que corresponden a tejidos y tipos celulares con diferentes patologías, o que derivan de lineas inmortalizadas, o han sido inmortalizadas, o están influenciadas a distintas condiciones experimentales como genes *knock-out* y mutaciones.

3.3.2. *ComBat* consigue una agrupación de todos los tipos celulares presentes en más de un GSE

Para la normalización de los datos para la construcción del paisaje de Waddington se seleccionaron 5 muestras máximo (si las hubiere) de cada tipo celular, con el objetivo de que los actuales

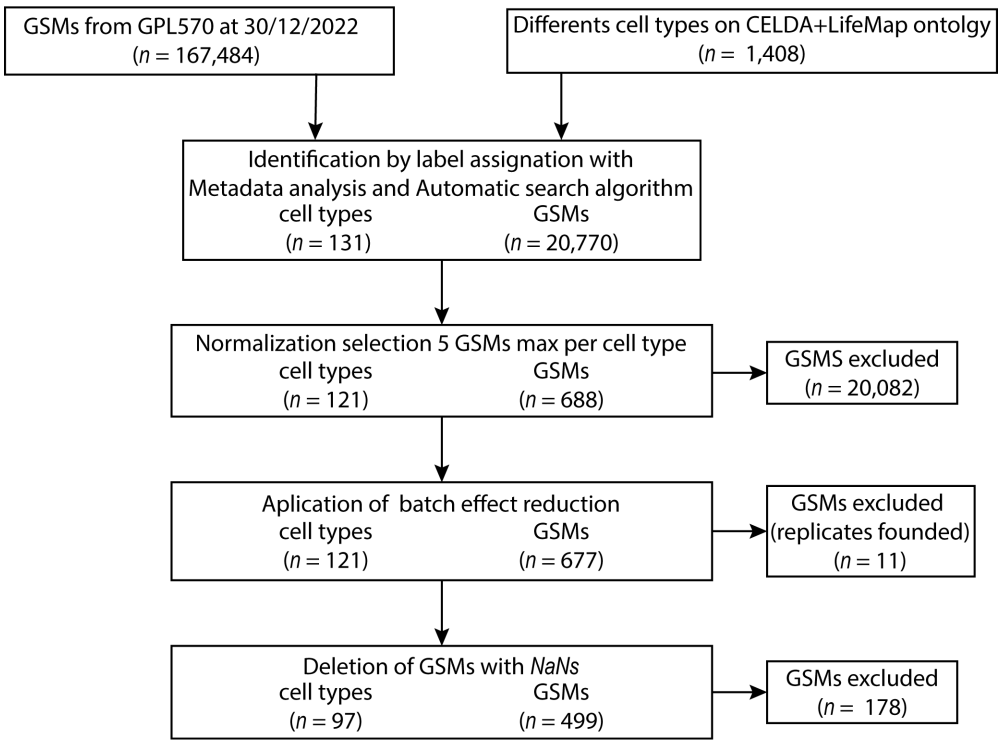


Figura 3.7: **Diagrama del total de muestras obtenidas y descartadas en cada uno de los pasos.** El diagrama cubre desde el origen de las muestras alojadas en GEO para la plataforma GPL570, pasando por la identificación de muestras, la normalización del conjunto de muestras para la construcción del paisaje de Waddington, la aplicación de mecanismos de reducción del *batch effect* y finalmente la eliminación de aquellas muestras con valores de genes nulos (*NaNs: Not a Number*)

algoritmos de normalización de RMA puedan soportar la carga de trabajo y no colapsaran provocando el fallo del programa por falta de memoria. La normalización detectó fallos en algunas de las muestras lo cual terminó por excluir un total de 10 tipos celulares de la normalización final, contando finalmente con un total de 121 tipos celulares y 688 muestras normalizadas conjuntamente (**Fig. 3.7**).

En cuanto a los GSEs de hematopoyesis se normalizaron conjuntamente sin la necesidad de tener que seleccionar un número de muestras por cada tipo (**Fig. 3.8**). Tras normalizar los datos transcriptómicos hematopoyéticos conjuntamente con el algoritmo RMA se realizó un PCA (**Fig. 3.9**) para observar y cuantificar la agrupación. Después de aplicar las técnicas de reducción del *batch effect* sobre el mismo conjunto de datos se volvió a realizar un nuevo PCA (**Fig. 3.9**). A partir de la distancia bidimensional de estos puntos utilizando las dos primeras componentes principales como las coordenadas x e y de cada punto se calcula la distancia antes y después de aplicar las técnicas de reducción de *batch effect*.

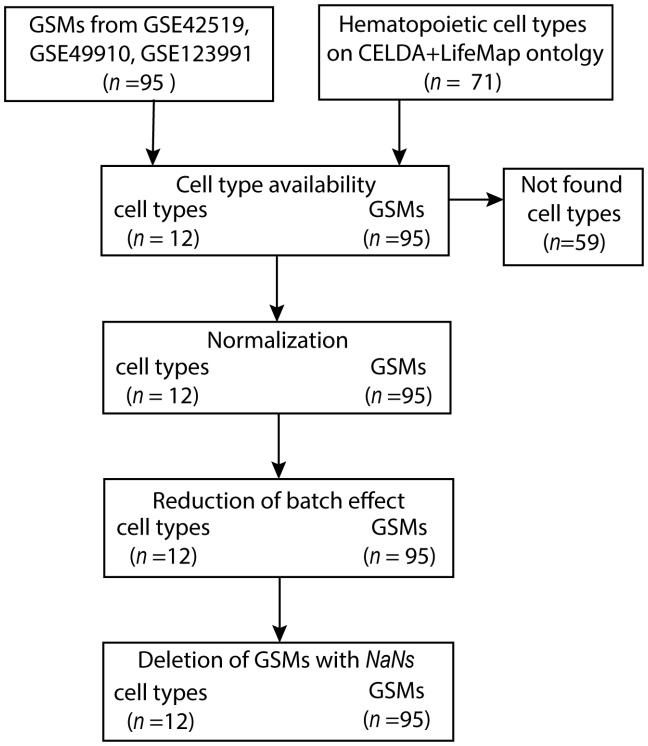


Figura 3.8: **Diagrama del total de muestras obtenidas y descartadas en cada uno de los pasos para las muestras seleccionadas para la Prueba de Concepto**.Partiendo inicialmente en 3 GSEs de datos de hematopoyesis (GSE42519, GSE49910, GSE123991). Pasando por la identificación de los datos alojados en los 3 GSEs, la normalización conjunta, la aplicación de mecanismos de reducción del *batch effect* y finalmente la eliminación de aquellas muestras con valores de genes nulos (*NaNs: Not a Number*)

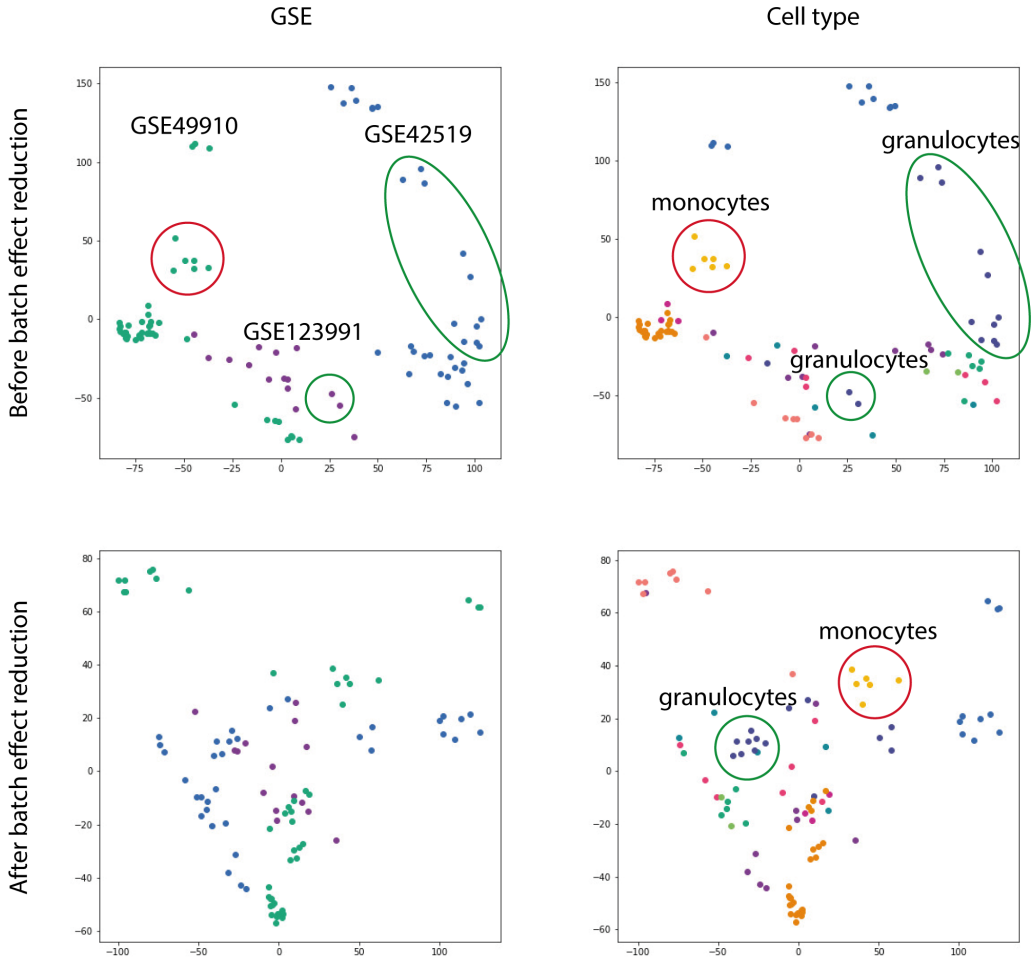


Figura 3.9: **PCAs antes y después de corregir el *batch effect* sobre los datos hematopoyéticos**. Se resaltan en un círculo rojo las muestras de monocitos y en un círculo verde las de granulocitos. Tras aplicar la corrección del *batch effect*, los monocitos que se hallaban únicamente en el GSE49910 permanecen igualmente agrupados, mientras que los granulocitos presentes en el GSE123991 y en el GSE42519 aparecen dispersos antes de la corrección del *batch effect* y agrupadas después, la medida de las distancias de estos y los otros tipos celulares se muestra en la **Tabla 3.2**.

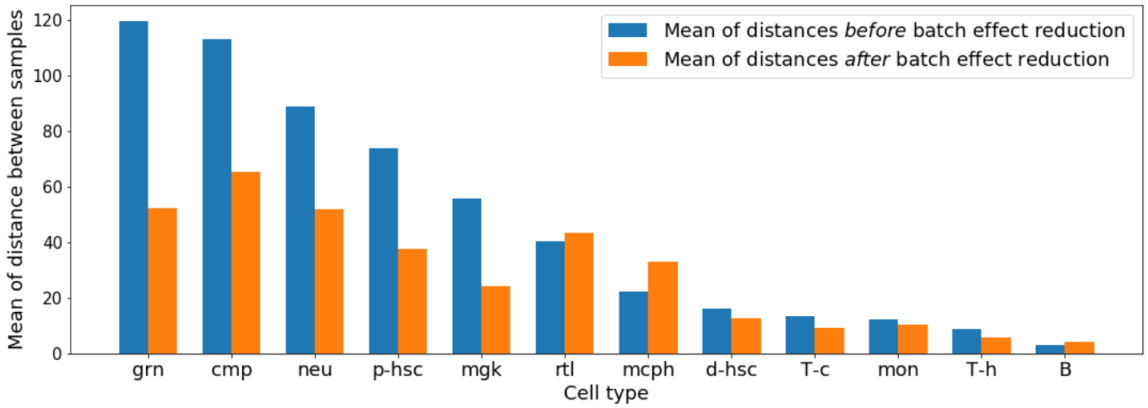


Figura 3.10: **Diagrama de barras de la distancia media de los tipos celulares con la corrección del *batch effect***. Se observa la distancia media entre los puntos de los tipos celulares antes (azul) y después (naranja) de aplicar la corrección del *batch effect* en todos los tipos celulares. Se utilizan los nombre simplificados para los tipos celulares (**Tabla 3.2**). Con un asterisco se indican aquellos tipos celulares presentes en más de un GSE

A simple vista, la Figs. **Fig. 3.9**, no clarifica, suficientemente el grado de agrupación de las muestras antes y después de aplicar técnicas de reducción del *batch effect*, es por ello que para poder evaluar la efectividad de estas técnicas se recurrió al cálculo de la media de las distancias entre puntos de una misma agrupación (tipo celular o GSE).

Los resultados de la media de las distancias se muestran en la **Tabla 3.2** donde se observa que aproximadamente el ∼70 % de los tipos celulares se agrupan más próximamente tras aplicar las técnicas de reducción del *batch effect* y que el 100 % de los tipos celulares que tienen al menos presencia en 2 GSEs se han agrupado tras aplicar técnicas de reducción del *batch effect*. Este experimento indica que *ComBat* es más efectivo cuantos más tipos celulares tengan muestras en diferentes GSEs y que *ComBat* ayuda a reducir el *batch effect* para casi todo el conjunto de datos utilizado.

En el caso de los GSEs se observa que las técnicas de reducción del *batch effect* han alejado 2 de los 3 conjuntos utilizados. En caso de que estas técnicas hubieran aproximado todos los GSEs (o al menos la mayoría, 2 de 3) se puede considerar que *ComBat* tiende a agrupar indistintamente todos los tipos celulares.

Atendiendo a las diferencias de distancias de la **Fig. 3.10**, la **Fig. 3.11** y en la **Tabla 3.2**, *ComBat* reduce de forma significativa la distancia media entre puntos para aquellos tipos celulares para los cuales hay muestras en más de un GSE.

Los tipos celulares que están presentes en un único GSE experimentan aproximadamente la mitad

		Means		
Name	**Simplified name**	**Before**	**After**	**Variation**
* common myeloid progenitor	cmp	113.00	65.53	better
* megakaryocyte erythroid precursor cells	mgk	55.67	24.15	better
t helper cells	T-h	8.90	5.62	better
* neutrophils	neu	89.03	51.78	better
monocytes	mon	12.33	10.41	better
* primitive hematopoietic stem cells	p-hsc	74.00	37.85	better
* granulocyte monocyte progenitor cells	grn	119.55	52.49	better
definitive hematopoietic stem cells	d-hsc	16.25	12.54	better
t cytotoxic cells	T-c	13.41	9.26	better
macrophage dendritic cell progenitors	mcph	22.26	33.26	worst
reticulocytes	rtl	40.32	43.33	worst
mature b cells	B	3.08	4.29	worst
GSE49910		101.00	132.04	worst
GSE42519		175.04	162.26	better
GSE123991		50.07	52.40	worst

Tabla. 3.2: **Resultados de las distancias medias anteriores y posteriores a la aplicación de técnicas de reducción del *batch effect*.** En la primera columna se muestran las diferentes agrupaciones de datos, 1 por cada tipo celular y una para cada GSE. La segunda columna corresponde a la media de las distancias antes de aplicar técnicas de reducción del *batch effect*. La tercera columna corresponde a la media de las distancias después de aplicar técnicas de reducción del *batch effect*. La cuarta columna corresponde a la diferencia de la media anterior y posterior, indicando *better* si la media ha disminuido tras aplicar técnicas de reducción del *batch effect* y *worst* si la media ha aumentado tras aplicar técnicas de reducción del *batch effect*. Con un asterisco se marcan aquellos tipos celulares que tienen presencia en más de un GSE.

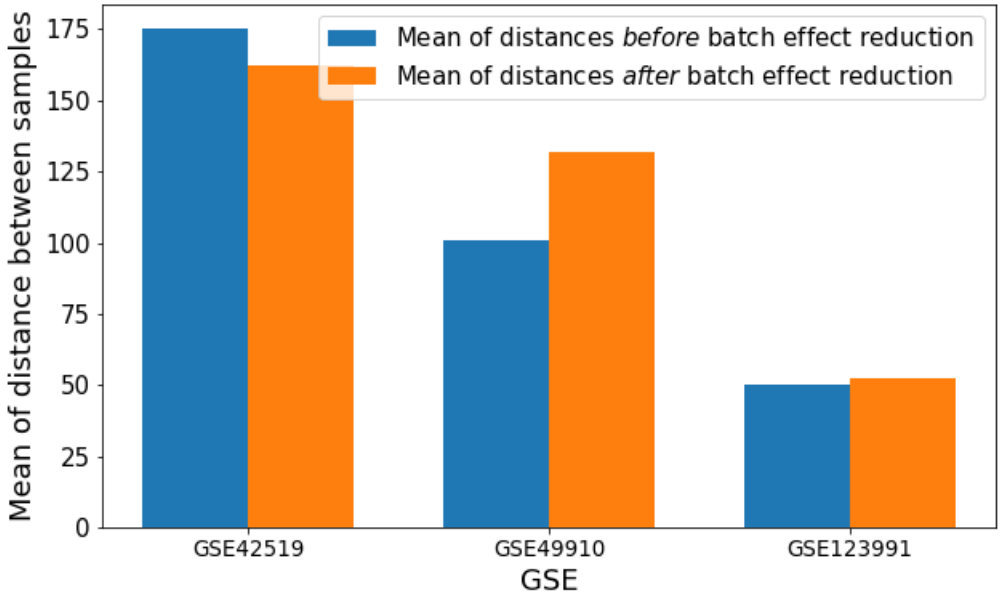


Figura 3.11: **Diagrama de barras de la distancia media en la corrección del *batch effect* de los GSEs**. Se observa la distancia media entre los puntos de los tipos celulares antes (azul) y después (naranja) de aplicar la corrección de *batch effect* para todos los GSEs

(4 de 7) una disminución de la distancia media entre ellos. En estos casos la diferencia entre la distancia media antes y después de aplicar técnicas de reducción del *batch effect* es menos acusada que en los casos en los que existe más de una muestra de ese tipo celular por GSE (**Fig. 3.10**, **Fig. 3.11**, y **Tabla 3.2**).

3.4. Discusión

Gracias al repositorio de GEO se dispone de una gran cantidad de datos transcriptómicos de diferentes estudios elaborados por diferentes grupos de investigación (Edgar et al., 2002). Sin embargo, la reutilización y reanálisis de estos datos no entra dentro de los objetivos del propio repositorio (Edgar et al., 2002).

Aun así, la posibilidad de poder reutilizar datos generados para otros análisis implica un gran ahorro tanto de tiempo como económico además de servir como una revalidación de los datos alojados. En este capítulo se han generado algoritmos para facilitar la reutilización de datos del repositorio de GEO, y también evaluar *ComBat* como herramienta para reducir el *batch effect*. Así, el *pipeline* aquí desarrollado incorpora dos nuevas tareas de procesamiento automático de datos, una de identificación y etiquetado de muestras y otra de normalización de las mismas para su

posterior análisis.

Respecto al etiquetado de muestras, las normas de GEO obligan que los diferentes investigadores a la hora de subir las muestras rellenen manualmente los diferentes campos de metadatos. De todas formas, aunque los campos de datos GEO son fijos, su contenido es flexible, permitiendo a los investigadores describir la naturaleza de las muestras, el trabajo, etc, con lenguaje natural no estandarizado y subjetivo (Edgar et al., 2002), lo cual dificulta en gran medida la reutilización automática de los mismos. Dada la gran cantidad de muestras alojadas en GEO y la forma de descripción de los metadatos plantea la posibilidad de crear algoritmos de interpretación de texto para la búsqueda de tipos celulares concretos (Algoritmo de Búsquedas Automatizadas) y para la interpretación de los metadatos (Algoritmo de análisis de metadatos).

Para cumplir nuestro objetivo de conseguir el mayor número posible de muestras etiquetadas se planteó atacar el problema desde dos prismas. El primero con el Algoritmo de Búsquedas Automatizadas, en el cual permite aprovechar el intérprete de texto del propio buscador de NCBI y el Algoritmo de análisis de metadatos el cual brinda la posibilidad de implementar nuestro propio intérprete basado en Levenshtein.

Es posible que ambos algoritmos utilizados para el etiquetado de muestras tengan cierto solapamiento entre ambos. Un estudio más exhaustivo sobre el tema, analizando más plataformas podría arrojar más luz sobre el grado de solapamiento de ambas técnicas y podría evaluar cual de ellas consigue un mayor número de muestras etiquetadas satisfactoriamente. Como nuestro objetivo con respecto al etiquetado de muestras es conseguir el mayor número posible de muestras transcriptómicas etiquetadas se decidió utilizar ambas estrategias simultáneamente centrándonos en una sola plataforma.

Una posible línea para el futuro podría ser la evaluación de las diferencias de etiquetado de ambas estrategias (el Algoritmo de Búsquedas Automatizadas y el Algoritmo de análisis de metadatos), siendo la más prometedora la estrategia del Análisis de Metadatos puesto que es donde podemos incidir más mejorando los interpretes de texto que utiliza el algoritmo. En esta línea, sería posible la implementación de herramientas más potentes de procesamiento de lenguaje natural Brown et al. (2020) en lugar de la métrica de similitud de Levenshtein Levenshtein (1966). Utilizar técnicas de procesamiento de lenguaje natural automáticas nos permitiría un análisis más profundo de los metadatos e incluso incluir los artículos científicos relativos a la muestra publicada lo cual se traduciría en una mejoría tanto cuantitativa como cualitativa del algoritmo.

Aunque el porcentaje de muestras etiquetadas con tipos celulares de la plataforma GPL570 no sea alto, es más que suficiente para realizar la construcción del Paisaje transcriptómico de Waddington, ya que, atendiendo a los resultados obtenidos, los datos están homogéneamente distri-

buidos a lo largo de los diferentes linajes y del dígrafo las muestras etiquetadas, permitiendo una extrapolación entre los tipos celulares para los cuales existen datos.

En cuanto a la normalización de las muestras en conjunto, se asume que existe un *batch effect* entre ellas, este es el principal motivo de utilizar una misma plataforma, la GPL570. Utilizar una combinación de plataformas diferentes, siempre del mismo organismo y analizando el mismo tipo de molécula, podría ocasionar un mayor *batch effect*. Queda abierta esta línea de investigación de la elaboración de algoritmos para el reanálisis de datos multi-plataforma incluyendo datos se secuenciación siguiente generación (NGS).

La normalización de los datos es un paso crucial y muy poco optimizado, un punto de mejora importante sería la readaptación de algoritmos de normalización para su procesamiento incremental y paralelización de cara a facilitar el procesamiento *Big data*. También es interesante establecer algún tipo de selección de un número máximo de muestras para evitar que un tipo celular esté sobre representado en la matriz final de datos. Pese a la pérdida de 10 tipos celulares, el total de tipos celulares etiquetados, su distribución y el número de muestras permite continuar en la elaboración de un Paisaje transcriptómico de Waddington.

Adicionalmente, a la luz de estos resultados se puede concluir que encontrar muestras localizadas en más de un GSE es importante para un mejor funcionamiento de *ComBat*, y que también puede minimizar el *batch effect* en la mayoría de muestras aunque estas solo estén presentes en un único GSE. Por tanto, *ComBat* resulta ser una herramienta que incorporar en nuestro *pipeline* como corrector del *batch effect* a utilizar en el los siguientes capítulos.

Capítulo 4

Construcción del paisaje de Waddington de hematopoyesis mediante algoritmos genéticos

4.1. Introducción

El siguiente paso necesario para la construcción de un paisaje de Waddington es el diseño de un algoritmo que nos permita maximizar la 'función de energía *quasi*-potencial'del paisaje de Waddington, es decir, elegir un grupo de genes que según nuestra hipótesis contribuyan en significativamente a su construcción.

El método escogido para ello hace uso de un algoritmo genético (Holland, 1992), el cual permite 'simular la selección natural' en un problema de optimización combinatoria. La función a optimizar en un algoritmo genético recibe el nombre de función objetivo. Esta función debe ser lo suficientemente restrictiva como para que el número de posibles candidatos de paisaje de Waddington no sea el genoma completo y por otro lado, lo suficientemente permisiva como para generar algún candidato a paisaje de Waddington válido.

Es necesario observar como responde la función objetivo a las diferentes combinaciones del conjunto de datos los mismos y definir los valores de los parámetros que rigen el funcionamiento del algoritmo genético, para así seleccionar aquellos valores de parámetros que mejor favorezcan el proceso de optimización (mejores puntuaciones máximas de la función objetivo, menor consumo de recursos, mayor velocidad de convergencia, etc).

Para realizar las pruebas y ajustes necesarios para establecer tanto la función objetivo como el protocolo de ajuste de parámetros es esencial la realización de una prueba de concepto. La prueba de concepto requiere un ajuste heurístico, lo cual consume recursos computacionales. Por ello es conveniente trabajar un con un subconjunto más pequeño de datos.

A lo largo de este capítulo vamos a desarrollar a nivel teórico el conjunto de elementos necesarios para el diseño del algoritmo genético. A continuación los pondremos a prueba con un conjunto de datos reducido mediante una prueba de concepto donde evaluaremos tanto el grado de convergencia y funcionalidad del algoritmo genético como los resultados obtenidos.

4.2. Materiales y métodos

4.2.1. Muestra y relaciones ontológicas

Como se mencionó previamente, para la construcción de la prueba de concepto se escogieron datos transcriptómicos relacionados con la hematopoyesis. El principal motivo es su disponibilidad y su amplia categorización (disponemos de información amplia sobre el desarrollo y la diferenciación de estos tipos celulares).

Partimos de los resultados obtenidos en los capítulos anteriores. Por un lado, tenemos una relación jerárquica ontológica de diferenciación celular fruto de la fusión mediante FOntCell de CELDA + LifeMap, utilizando el método de alineamiento coseno ya que según los valores de *F-score* evaluados en dicho capítulo, el método coseno presenta un mayor $F_{0,5}$ lo que nos indica que es aquel que resulta más preciso y por tanto el que menos errores introduce. De la ontología resultante extraeremos el subconjunto referente al desarrollo hematopoyético.

Por otro lado, se van a seleccionar manualmente muestras de datos transcriptómicos de GEO GPL570 de tipos celulares referentes a hematopoyesis. Concretamente hemos seleccionado 3 GSEs de datos de hematopoyesis: GSE42519, GSE49910, GSE123991, los cuales corresponden a un total de 95 muestras diferentes, y cubren un total de 12 tipos celulares (Mabbott et al., 2013);(Rapin et al., 2014).

4.2.2. Algoritmo genético

El objetivo del algoritmo genético es la optimización heurística de la función objetivo para seleccionar un conjunto de genes que representen la diferenciación celular.

La razón de utilizar un algoritmo genético reside en su capacidad de encontrar los conjuntos de datos que optimizan una función sin tener que probar todas las combinaciones, lo cual es compu-

tacionalmente prohibitivo dada la naturaleza *NP-hard* del problema. Por ello, teniendo en cuenta el volumen de datos que estamos manejando y sin conocer con exactitud el tamaño óptimo de cada conjunto, nos encontramos con un problema de combinatoria que no es manejable a nivel computacional. El problema aumenta cuando queremos modificar y probar diferentes tamaños de conjuntos de genes, por tanto, es inabarcable de resolver a fuerza bruta. La ecuación que describe el número de combinaciones posibles es:

$$C\binom{n}{r} = \frac{n!}{r! \cdot (n-r)!} \qquad \text{(Ec. 4.2.1)}$$

Donde C corresponde al total de combinaciones posibles. n representa el total de datos y r el total de conjuntos de datos. Por ejemplo, si utilizáramos conjuntos de $r = 50\ genes$, para poder evaluar todas las combinaciones posibles de un total de $n = 23500\ genes$ anotados en la GPL570 tendríamos:

$$C\binom{23500}{50} = \frac{23500!}{50! \cdot (23500-50)!} = 1,12 \cdot 10^{154}\ combinaciones \qquad \text{(Ec. 4.2.2)}$$

Con lo que se tendría más combinatorias posibles para conjuntos de 50 que átomos hay en el universo, y sumado al hecho de que queremos probar diversos conjuntos de genes puede dificultar considerablemente la tarea a nivel computacional. Es por ello, que es necesario utilizar un proceso que nos permita seleccionar los genes que mejor se ajusten a nuestro modelo sin tener que probar todas las combinaciones posibles.

Entendemos como algoritmo genético un tipo de algoritmo inspirado en el modelo de selección natural y la evolución. Los algoritmos genéticos son herramientas de optimización que actúan sobre un conjunto de datos a los que se les aplica una función objetivo la cual debe optimizarse. Los algoritmos genéticos nos permiten la optimización de funciones y a la vez la conservación de la información crítica de los datos (Whitley, 1994); (Holland, 1992).

De acuerdo con la **Eq. 4.2.3 a** el algoritmo genético es un proceso por el cual maximizamos una función $f(vc, td)$. En nuestra instanciación para el caso de la construcción del paisaje de Waddington x correspondería a las variables a optimizar vc, td (**Eq. 4.2.3 a**) y donde intervienen un conjunto de parámetros (**Eq. 4.2.3 c**). Este proceso se muestra esquematizado en **Fig. 4.1**.

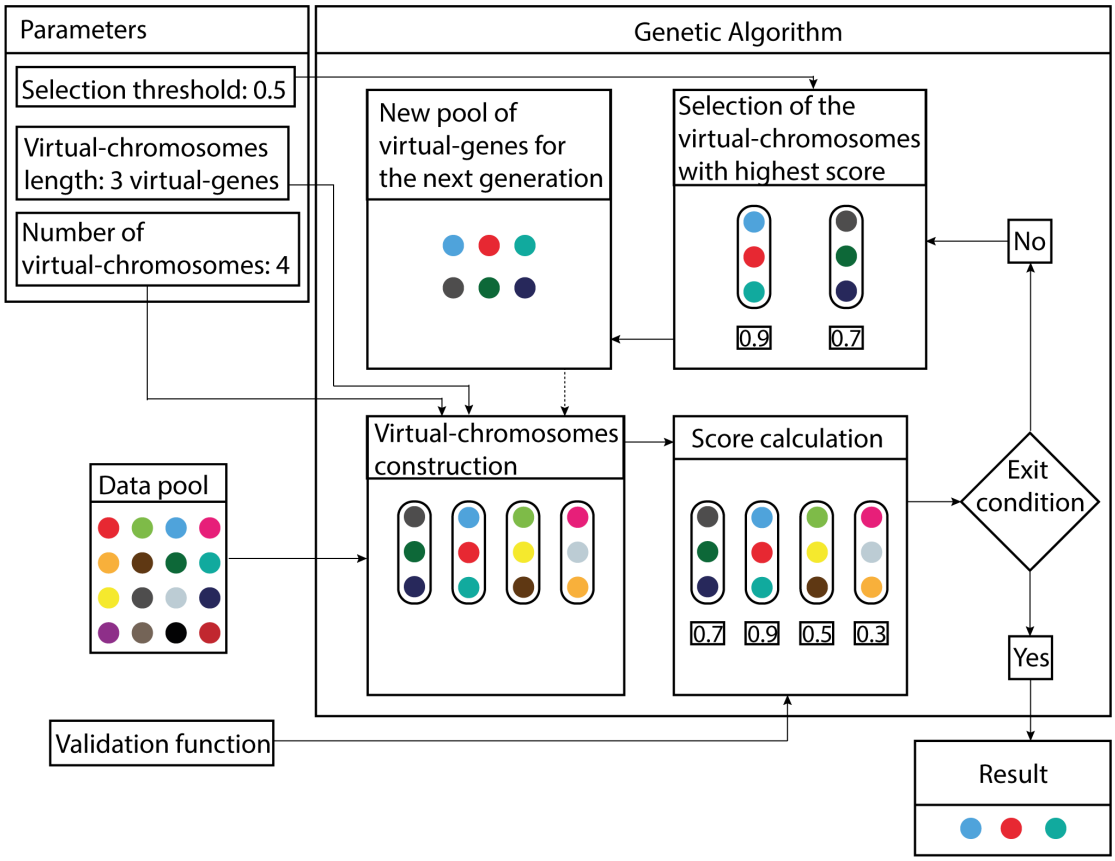


Figura 4.1: **Diagrama del funcionamiento básico del Algoritmo Genético.** Consta de 5 'elementos' encuadrados. El primero, etiquetado como *Parameters* corresponde al conjunto de parámetros que del algoritmo y definen su funcionamiento. El segundo, etiquetado como *Data pool*, corresponde al conjunto de datos introducidos en el algoritmo, en nuestro caso información transcriptómica de los genes. Una vez dentro del algoritmo estos genes pasarán a ser los genes-virtuales. El tercero, etiquetado como *Fitness function* calcula las puntuaciones de los cromosomas-virtuales con el objetivo de seleccionar los conjuntos de genes-virtuales que más se adapten. El cuarto, etiquetado como *Genetic algorithm*, representa el funcionamiento del algoritmo genético secuenciado en el bucle de: construcción de cromosomas-virtuales a partir de los genes-virtuales, evaluación de cada cromosoma-virtual, evaluación de la condición de salida, selección de aquellos con mayor puntuación y cerrando el ciclo con la generación de un nuevo *pool* de genes. El quinto y último elemento es el resultado, etiquetado como *Result*, el cual ofrece el conjunto de genes-virtuales del cromosoma-virtual con mayor puntuación del último ciclo (aquel que satisfizo la condición de salida).

$$
\begin{aligned}
&a)\ \text{máx}\, f(vc, td);\\
&b)\ vc = \{vg\};\\
&c)\ parameters = \{L_{vc},\ N_{vc},\ O,\ N_{gr},\ \theta\};\\
&d)\ L_{vc} = length(vc)\\
&e)\ N_{vc} = \#vc\\
&f)\ N_{gr} = \#gr\\
&g)\ \theta = (\theta_p,\ \theta_s)\\
&h)\ O = \{ct, rel\};\\
&i)\ EC = \begin{cases} True & if\ \ max(vc_j) >= \theta_s \\ True & if\ \ gr = N_{gr} \\ True & if\ \ otherwise \end{cases}
\end{aligned}
\qquad \text{(Ec. 4.2.3)}
$$

Para evitar confusiones nos referiremos a los cromosomas del algoritmo genético como 'cromosomas-virtuales' o siguiendo la notación como vc, diferenciándolos de los cromosomas biológicos; y paralelamente nos referiremos a los genes del algoritmo genético como genes-virtuales o siguiendo la notación como vg.

Las variables a optimizar en la instanciación del algoritmo genético para la construcción del paisaje de Waddington son:

- vc: Los cromosomas-virtuales. Los vc se definen como un conjunto de genes-virtuales, vg (**Eq. 4.2.3 b**). Los vg por su parte son el nombre de cada gen, por lo que nos permite acceder a la expresión transcriptómica td de un gen para el conjunto de tipos celulares. Los vc son cada una de las soluciones generadas por el algoritmo para la optimización de la función objetivo $f(vc, td)$.
- td: es una matriz de expresión transcriptómica relacionada de los tipos celulares ct y genes. La expresión transcriptómica de cada gen irá codificada con el nombre del gen en la variable vg.

El conjunto de parámetros de **Eq. 4.2.3 c** son:

- L_{vc}: Tamaño de los cromosomas-virtuales (vc), es decir el número de vg que tiene cada vc (**Eq. 4.2.3 d**).
- N_{vc}: Número de cromosomas-virtuales. Es el tamaño de la población del algoritmo genético,

es decir, es el total de vc que intervienen en el algoritmo genético (**Eq. 4.2.3 e**).

- N_{gr}: Número de generaciones, siendo una generación (gr) cada iteración del el ciclo de construcción de la población de soluciones N_{vc}, evaluación de los vc y finalmente selección de los vc con la puntuación más alta (**Eq. 4.2.3 f**).
- θ: Es un vector de umbrales de selección (θ_p, θ_s). θ_p definida como el percentil de vc que se desechan después de que hayan sido evaluados por la función objetivo $f(vc, td)$ y se les haya asignado una puntuación. θ_s es el umbral de puntuación que una vez alcanzado por un vc_j satisfaría la condición de salida $ExitCondition$ y pondría fin al algoritmo genético presentando ese vc_j como resultado (**Eq. 4.2.3 g**).
- O: Representa la ontología de las muestras respecto a la jerarquía de diferenciación. O se define (**Eq. 4.2.3 h**) también como un conjunto compuesto de: ct lista de tipos celulares que van a ser evaluados por el algoritmo genético y rel que es la jerarquía de diferenciación entre los diferentes ct que viene dada como una matriz de dos columnas donde en la primera columna se representan los antecesores y en la segunda uno de los descendientes.
- EC: es el parámetro que nos indica la condición de salida (*Exit Condition*), definida como una puntuación de un vc_j que supere el valor θ_s o bien la superación de un N_{gr} determinado de iteraciones (**Eq. 4.2.3 i**).

En el algoritmo genético (**Fig. 4.1**) partimos de los vg que corresponden a uno de los genes del conjunto de muestras de la plataforma GPL570, los cuales constituyen un vector de la transcripción de ese gen en cada uno de los tipos celulares. A partir de estos vg se generan los vc (**Eq. 4.2.3 b**). Cada vc tendrá el mismo número determinado de vg, seleccionados de forma aleatoria de la totalidad de vg disponibles.

En cada generación gr, se estimará el valor de la función objetivo de cada solución vc. $f(vc, td)$ recibe la información contenida en la ontología O. Las soluciones vg son evaluados por $f(vc, td)$ como un conjunto (un vc).

Con la evaluación de cada solución vc se elegirá un porcentaje de ellos a partir de θ_p, en nuestro caso fijado en 0,5. Aquellas soluciones vc seleccionadas continuarán siendo procesados en la siguiente generación por el algoritmo genético y el resto será desechado (Fig. 4.1).

Los vc seleccionados son aquellos que han obtenido las puntuaciones más altas por la función objetivo $f(vc, td)$. Esto constituye el final de un ciclo de generación. Si se cumple la condición de parada EC del algoritmo genético, este termina proporcionando como resultado la solución vc con mayor puntuación, en caso contrario se continúa con los ciclos de generación de una nueva población de soluciones vc.

En el caso que se deba crear una nueva generación, las nuevas soluciones vc se construirán a partir del conjunto de vg de las soluciones vc seleccionados en la anterior iteración gr, es decir, el percentil superior designado con θ_p. Mediante este proceso después de sucesivas generaciones se esperan puntuaciones más elevadas en las soluciones vc y se seleccionen aquellos vg que optimicen la función objetivo $f(vc, td)$.

El criterio de parada EC puede satisfacerse tanto si un vc supera θ_s como también si se cumple la segunda condición y se supere el N_{gr}. Esta segunda opción es explorada para la realización de los estudios de convergencia y asegurar que se han alcanzado los máximos generales de la función objetivo. En cualquiera de los dos casos, el resultado sería el conjunto de vg del vc con mayor puntuación. Este conjunto de vg es un conjunto de genes cuyos valores de expresión transcriptómica satisfacen en mayor medida $f(vc, td)$, establecida para la construcción del paisaje de Waddington para el conjunto de ct introducidos.

En resumen, el algoritmo genético va evaluando los diferentes vc y seleccionado aquellos que obtengan una mayor puntuación de $f(vc, td)$. Posteriormente, se genera un nuevo conjunto de vg que reconstruirán unos nuevos vc para ser evaluados en la siguiente generación hasta satisfacer EC, que en nuestro caso es el cumplir con el número máximo de generaciones N_{gr}.

4.2.3. Función objetivo

La función objetivo ($f(vc, td)$ **Eq. (4.2.3) a**) permite evaluar cada solución vc de modo que podamos seleccionar los vg a lo largo de las generaciones. La $f(vc, td)$ debe ser una aproximación computacional del modelo del paisaje de Waddington que queremos crear. En nuestro caso, buscamos un conjunto de genes cuya expresión vaya decreciendo conforme avanza el proceso de diferenciación, es decir, un conjunto de genes que 'expliquen' el desarrollo y diferenciación celular. Por tanto, la función objetivo $f(vc, td)$ debe tener en cuenta todas las relaciones de diferenciación que tienen los diferentes tipos celulares entre sí, es decir, debemos acudir a las relaciones ontológicas de ascendencia-descendencia generadas por FOntCell fruto de la fusión de las ontologías de LifeMap y CELDA (Cabau-Laporta et al., 2021).

Por otro lado, $f(vc, td)$ requiere los datos de transcripción td de cada gen en cada tipo celular ct. Para ello, previamente se realizará la media de expresión de cada ct para el conjunto de muestras que disponemos. También para estimar un 'Estatus transcriptómico' St definido como:

$$St_i = \sum_{i=1}^{ct} vc_j(td) \qquad \text{(Ec. 4.2.4)}$$

Donde para cada tipo celular i se le asignará un valor St calculando el sumatorio de la td correspondiente del conjunto de vg que se estén evaluando en el correspondiente vc.

Para construir la 'puntuación cromosómica' $score$ (la puntuación de cada solución vc) la función objetivo $f(vc, td)$ calcula el valor St (**Eq. (4.2.4)**) y es a partir de este en el que calculará el $score$ correspondiente dependiendo del método de la función objetivo $f(vc, td)$ que utilicemos. Para ello, hemos implementado tres funciones objetivo $f(vc, td)$ diferentes: decrecimiento básico (DB), decrecimiento creódico (DC) y decaimiento exponencial creódico (DEC).

4.2.3.1. Función objetivo: decrecimiento básico (DB)

A modo de prototipo inicial se introdujo el método por decrecimiento básico (DB) para definir la función objetivo $f(vc, td)$. Este método genera el estatus transcriptómico St de cada tipo celular. A continuación, para calcular el $score$ de cada solución vc se utiliza:

$$f(vc_j, td) = \frac{rel(decrecientes)}{rel} \qquad \text{(Ec. 4.2.5)}$$

$$rel(decrecientes) = \begin{cases} 1 & if\ \ St_i >= St_{i+1} \\ 0 & if\ \ otherwise \end{cases} \qquad \text{(Ec. 4.2.6)}$$

donde $rel(decrecientes)$ consiste en el conjunto de relaciones rel cuyo estatus transcriptómico St_i del tipo celular ct_i ascendente es mayor que el St_{i+1} (**Eq. (4.2.4)**) del ct_{i+1} descendiente. A continuación, se calcula el porcentaje de $rel(decrecientes)$ respecto al total de relaciones rel. Este porcentaje es la puntuación del cromosoma-virtual vc. Por tanto el algoritmo genético irá seleccionando aquellos genes-virtuales vg que hagan el mayor número posible de $rel(decrecientes)$.

4.2.3.2. Función objetivo: decrecimiento creódico (DC)

El método de decrecimiento creódico (DC) tiene en cuenta la jerarquía de diferenciación rel ascendentes entre los tipos celulares ct, es decir, tiene en cuenta la información del creodo. Para cada tipo celular extrae la información relativa a su creodo a partir de la información de la ontología O:

$$\begin{aligned} &creode \rightarrow O; \\ &creode = \{rel, ct\} \end{aligned} \qquad \text{(Ec. 4.2.7)}$$

Nos interesan los tipos celulares ascendentes respecto al tipo celular i. Es por ello que definimos

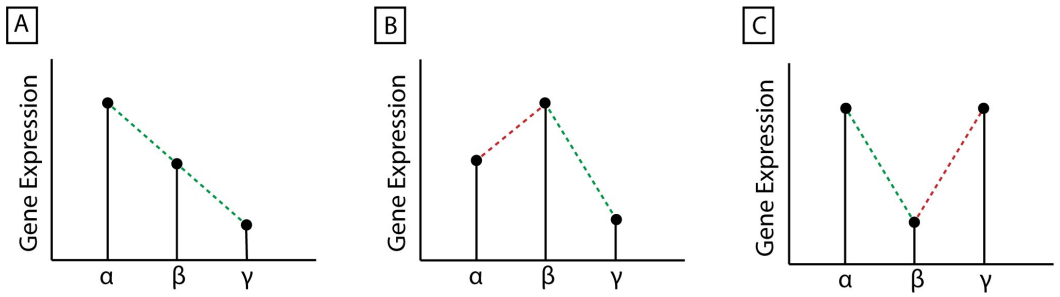


Figura 4.2: **Esquema del cálculo de las puntuaciones celulares del método de DC.** El orden de diferenciación de los tipos celulares en los tres creodos serían α, β, γ. En el creodo A observamos un decrecimiento constante por tanto la expresión transcriptómica de todos esos tipos celulares se sumaría. En el creodo B la expresión de α es menor que la de β y mayor que la de γ. Por lo tanto la expresión transcriptómica de α y γ se sumaría y al resultado se le restaría la expresión transcriptómica de β. En el creodo C la expresión transcriptómica de α sería mayor que la de β y menor que la de γ por lo tanto se sumaría la expresión transcriptómica de α y β y restaría la de γ. Obteniendo la puntuación de los tres creodos el resultado final sería la suma de la Puntuación de A + la Puntuación de B + la Puntuación de C.

la información creódica como $creode$. Este es un conjunto de información relativa a las jerarquías de diferenciación rel, los tipos celulares ct y los datos transcriptómicos td tanto de un tipo celular i como de los otros ct corriente arriba en el creodo $i + n$, siendo n el número de antecesores corriente arriba que evaluamos (por defecto $n = 3$).

La función objetivo $f(vc_j, td)$, parte en primer lugar del cálculo del estatus transcriptómico St (**Eq. (4.2.4)**) de cada uno de los tipos celulares ct. Para a continuación, teniendo en cuenta la información de $creode$, se calcula calcular la puntuación de cada tipo celular i mediante:

$$score_i^j = \sum_{i=1}^{n} St_i + \begin{cases} +St_{i+n} & if \;\; St_{i+n} > St_i \\ -St_{i+n} & if \;\; otherwise \end{cases} \qquad \text{(Ec. 4.2.8)}$$

donde $score_i$ se calcula sumando al estatus transcriptómico del tipo celular i (St_i, **Eq. (4.2.4)**) el estatus transcriptómico de los siguientes n antecesores (St_{i+n}) siempre y cuando cumpla la condición que St de cada tipo celular ascendente sea mayor que St_i. En caso contrario se computaría como un valor negativo el St_{i+n}, favoreciendo así las relaciones descendentes a lo largo de los creodos. Posteriormente, mediante el sumatorio de las $St_i + / - St_{i+n}$ se obtiene la $score_i$. Este proceso se muestra esquematizado en la **Fig. 4.2**.

$$score\ f(vc_j, td) = \sum_{i=1}^{ct} score_i^j \qquad \text{(Ec. 4.2.9)}$$

Finalmente, la suma de las $score_i$ será la puntuación final de la solución vc **Eq. (4.2.9)**. El algoritmo genético seleccionará aquellos vc que tengan valores de puntuación general más altos, ya que indican una mayor presencia y peso del tipo de relaciones ascendente-descendente que pretendemos que el algoritmo genético seleccione.

4.2.3.3. Función objetivo: decaimiento exponencial creódico (DEC)

El método DEC utiliza la información de cada creodo por separado y se basa en un hipotético decaimiento exponencial de la expresión de los genes a lo largo de cada creodo, durante la diferenciación. En este método se extrae de la ontología O la información relativa a los creodos **Eq. (4.2.7)**. A diferencia de las funciones anteriores, la función DEC computa los $creode$ teniendo en cuenta la totalidad de los tipos celulares ct. Los $creode$ son evaluados de forma paralela en el algoritmo genético, por lo que los resultados finales se expresan para cada uno de los creodos que intervengan en el conjunto de datos.

Por tanto, cada $creode$ dispone de sus propios vc evaluados de forma independiente. Cuando una solución vc es evaluada por este método, en primer lugar se calcula el estatus transcriptómico St **Eq. (4.2.4)** de los diferentes tipos celulares ct que intervienen en ese creodo. Después, alineados, se genera un ajuste a una curva exponencial mediante una ecuación con expresión $creodo = a_0 + a_1 \ln(x)$ donde x es un valor numérico asignado a cada tipo celular ct indicando su posición en el creodo de acuerdo a las relaciones de diferenciación rel e y es el estatus transcriptómico St de cada uno de los tipos celulares ct (**Fig. 4.3**) a_0 y a_1 son parámetros generados por las siguientes regresiones lineales.

$$\begin{aligned} &R^2 \rightarrow St_i = a_0 + a_1 \ln(rel_i); \\ &m \rightarrow St_i = a_0 + a_1 \ln(rel_i) \end{aligned} \qquad \text{(Ec. 4.2.10)}$$

Donde R^2 son los residuos cuadrados y m es la pendiente de la curva de ajuste. **Eq. (4.2.10)**.

$$f(vc_j, td) = score^j = \begin{cases} 0 & if\ m \geq 0 \\ R^2 & if\ otherwise \end{cases} \qquad \text{(Ec. 4.2.11)}$$

Se establece el $score^j$ para cada cromosoma-virtual vc_j dependiendo, en primer lugar, de si m es positivo o igual a 0, lo que implicaría $score_i = 0$ dado que esto representaría una curva ascendente (**Fig. 4.3**, A). Si el valor de m resultara negativo, estaríamos ante una curva decreciente, por lo que el propio ajuste de los residuos cuadrados R^2 es el valor que se asignaría a la $score^j$ y esta se asignaría al vc. Esto nos garantiza que el algoritmo genético va a propiciar la selección de aquellos

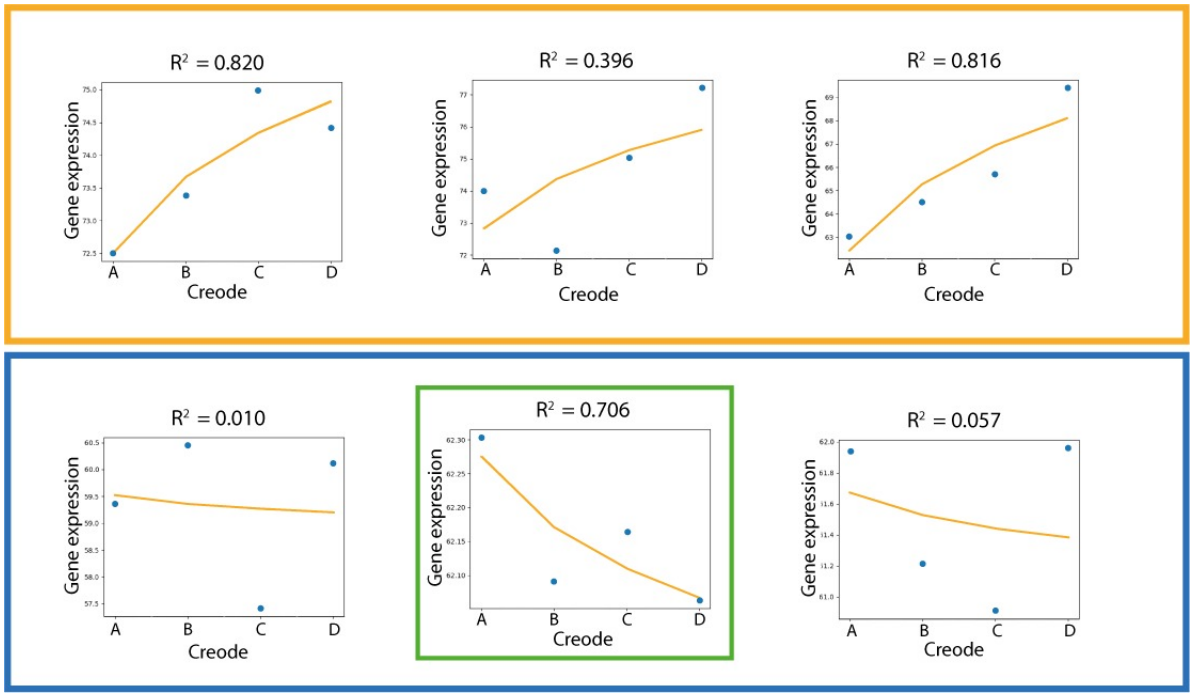


Figura 4.3: **Esquema del funcionamiento del método de validación por DEC**. Cada una de las gráficas representa la 'nube de puntos' (en azul) generada al representar el sumatorio de la expresión genética de un creodo para un cromosoma-virtual diferente. En este ejemplo el creodo estaría formado por los hipotéticos tipos celulares A, B, C y D, en orden del menos diferenciado al atractor. En naranja se representa la curva que de ajuste mediante regresión exponencial. En el recuadro naranja se observan aquellos valores que serían asignados con una puntuación de 0 al presentar una curva con una pendiente positiva (creciente). En el recuadro azul se muestran el resultado de los cromosomas-virtuales que al generar una curva con pendiente negativa (decreciente) se les asignaría el ajuste de los residuos cuadrados como su puntuación. Finalmente en el recuadro verde se resalta la gráfica con la mejor puntuación ya que tiene un mayor ajuste atendiendo a los residuos cuadrados además de tener una pendiente negativa.

conjuntos de vg cuya td sea exponencialmente decreciente a lo largo de los creodos.

4.2.4. Optimización de parámetros

4.2.4.1. Prototipo del funcionamiento del algoritmo genético

El objetivo de la creación de un prototipo es observar si el diseño del algoritmo consigue la convergencia de la función objetivo a lo largo de las generaciones, es decir, se pretende evaluar la propia lógica del algoritmo en un entorno incluso más sencillo que la propia prueba de concepto. Entendemos por convergencia de la función objetivo como el estado alcanzado por el algoritmo genético en el cual no se obtienen mejores resultados en las siguientes generaciones, entendiendo que se ha alcanzado un máximo global.

La lógica del algoritmo genético nos presenta principalmente 3 parámetros que podrían influir directamente en la convergencia del algoritmo genético y por tanto nos dan una idea del peso que tienen para la optimización de la función objetivo. Estos parámetros son el número de genes-virtuales vg, el número de cromosomas-virtuales vc y el número de generaciones N_{gr}.

Utilizando los GSEs de hematopoyesis seleccionados para la prueba de concepto se ejecutó una serie de pruebas con el algoritmo evaluando por separado cada uno de los parámetros de genes-virtuales vg, cromosomas-virtuales vc y número de generaciones N_{gr}, manteniendo el resto fijas para observar la influencia de cada una de ellas en la convergencia del algoritmo, ejecutando versiones paralelas con diferentes valores del parámetro de estudio. Por defecto los parámetros se fijarán como: número de genes-virtuales vg = 5, número de cromosomas-virtuales vc = 500, número de generaciones N_{gr} = 500, siempre y cuando en estas pruebas no se esté estudiando el efecto de ese parámetro en la convergencia.

Para la prueba del prototipado se utilizará la totalidad de los genes de la plataforma GPL570 de *Affymetrix*, un total de 23520 genes.

4.2.4.2. Relación entre variables

Podemos extraer una aproximación de la cantidad de generaciones necesarias para una convergencia que podrá ser recalculada en función de los resultados de convergencia observados en la prueba de concepto.

Contextualizando, los genes-virtuales vg y los cromosomas-virtuales vc en la plataforma GPL570 de *Affymetrix*, nos indica que las muestras poseen un total de genes, los cuales queremos que estén representados en la ejecución del algoritmo genético. Podemos calcular la probabilidad de que un gen tenga presencia o no en alguna de los cromosomas-virtuales vc partiendo del total de

genes que tenemos (*pool* de datos), los genes-virtuales por cromosoma-virtual (L_{vc}) y la cantidad de cromosomas-virtuales (N_{vc}), sabiendo que la construcción de los cromosomas-virtuales es aleatoria.

Establecemos la siguiente relación:

$$Probabilidad = \frac{1}{Ngenes} \cdot N_{vc} \cdot L_{vc} \qquad \text{(Ec. 4.2.12)}$$

donde la probabilidad de presencia de un gen es igual a 1 dividido por el número total de genes (en nuestro caso, $Ngenes = 23500$ por utilizar la plataforma GPL570 de *Affymetrix*) por el número de cromosomas-virtuales (N_{vc}) por el número de genes-virtuales (longitud de cromosomas-virtuales, L_{vc}).

En la prueba de concepto y futuras implementaciones de las que se quiera obtener información biológica tiene que darse una probabilidad que impliquen al menos que cada gen aparezca como mínimo con 1 repetición en alguno de los cromosomas virtuales vc del algoritmo genético. Para que todos los genes de los que disponemos puedan ser evaluados y no estemos perdiendo información cada vez que ejecutemos el algoritmo genético.

4.2.4.3. Optimización decrecimiento y decaimiento creódico

Esta optimización compara las diferencias de resultados y convergencia de ambas funciones objetivo y entender mejor el funcionamiento de diferentes combinaciones de parámetros de cara a escalar el algoritmo a una prueba con datos integrales para la construcción del paisaje de Waddington.

Para la optimización de parámetros se va ejecutar el algoritmo genético para cada una de las dos principales funciones objetivo que hemos diseñado (DC y DEC). Para ambos casos las combinaciones de parámetros se observan en la **Tabla 4.1**.

En ambos casos se parte de un total de 23520 genes del cual se va a eliminar el 60 % de los genes menos variables para quedarnos con un total de 9408. Esto aligerará la carga computacional del algoritmo ya que aquellos genes menos variables deberían ser depurados a lo largo de las generaciones y supondrían la introducción de ruido en el algoritmo. Para detectar los genes que tiene una mayor variabilidad en las diferentes muestras se generó un vector de la expresión para cada gen a lo largo de las muestras y se calculó la desviación estándar (Altman, Bland, 1995) de cada uno de estos vectores. Después se genera una lista de genes y su correspondiente desviación estándar. Se ordenan de mayor a menor desviación estándar y se extraen el 40 % con mayor

Tipo de test	**Análisis de la convergencia del algoritmo genético**										**Índice de estabilidad**	
Función objetivo	**DC**					**DC/DEC**					**DC**	**DEC**
#genes (*data pool*)	23520					9408					9408	9408
L_{vc}	10	20	50	100	150	10	20	50	100	150	150	150
N_{vc}	2350	1200	450	250	150	950	450	200	100	50	100	100
N_{gr}	500	500	500	500	500	500	500	500	500	500	200	200
θ_p	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
θ_s	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla. 4.1: **Tabla de parámetros de las pruebas del algoritmo genético.** En esta tabla se muestran las diferentes combinaciones de parámetros de las diferentes pruebas ejecutadas para la validación del algoritmo genético.

deviacisón estándar.

4.2.5. Análisis de enriquecimiento de ontología de genes (GO)

Existen ontologías de genes codificantes de proteínas. En estas ontologías cada clase representa a un gen de un organismo. Estas clases tienen una serie de términos que describen las funciones de los genes, la clase de proteína que codifican, además de otra información relevante a estos genes (Mi et al., 2012).

Un análisis de enriquecimiento de ontologías de genes parte de un conjunto de genes identificados por la ID de la ontología (u otro tipo si la herramienta lo permite) y mediante el acceso a una ontología de genes calcula el 'enriquecimiento estadísticamente significativo' de los términos de la ontología, es decir, busca si dentro del conjunto de genes que queremos analizar hay un sunconjunto que comparten una función que está en una proporción mayor (enriquecida) a la muestra general de referencia (la totalidad del genoma humano, por ejemplo) de modo que resulte ser estadísticamente significativa (Mi et al., 2012).

El algoritmo genético genera como resultado un conjunto de genes fruto de la optimización de la función objetivo a lo largo de un número de generaciones N_{gr}. Un análisis de GO permite saber si existe un enriquecimiento en la función de los genes seleccionados.

Para este análisis utilizaremos la herramienta PANTHER, la cual dispone de una base de datos de genes codificantes de proteínas con sus correspondientes funciones además de la subfamilia que pertenecen y la clase de proteína que codifican (Mi et al., 2012).

Para el análisis de enriquecimiento utilizaremos los resultados del DC y del DEC, para las distintas combinaciones de genes-virtuales y cromosomas-virtuales. Para el DEC se agregarán los genes-virtuales resultantes para cada uno de los creodos, mientras que para el DC se usarán los genes-virtuales del cromosoma-virtual con mayor puntuación en la última generación, ya que el método

Tipo de test	**Construcción paisaje de Waddington Inmunológico**										**Construcción paisaje de Waddington integral**		
Función objetivo	**DC**					**DEC**					**DEC**		
#genes (*data pool*)	9408					9409					9408		
L_{vc}	10	20	50	100	150	10	20	50	100	150	50	75	100
N_{vc}	950	450	200	100	50	950	450	200	100	50	200	150	100
N_{gr}	500	500	500	500	500	500	500	500	500	500	500	500	500
θ_p	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
θ_s	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla. 4.2: **Tabla de parámetros del algoritmo genético para la construcción del paisaje de Waddington de datos inmunológicos y del paisaje de Waddington integral.** En esta tabla se muestran los parámetros y sus valores del algoritmo genético para la construcción del paisaje de Waddington de datos inmunológico y para la construcción del paisaje de Waddington integral.

por DC de la función objetivo no separa los resultados por sus distintos creodos. Los parámetros utilizados para la construcción del paisaje de Waddington con el que extraeremos los genes para realizar el GO se muestran en la **Tabla 4.2**

4.3. Resultados

Para evaluar la convergencia del algoritmo genético representamos la media de las puntuaciones de los cromosomas-virtuales seleccionados con el paso de las generaciones. El funcionamiento del algoritmo genético nos representa un aumento de la puntuación media de la población con el paso de las generaciones hasta un estancamiento. En el funcionamiento de un algoritmo genético una vez alcanzado este estancamiento el comportamiento es oscilante. Si introducimos mucho ruido de forma accidental (por ejemplo genes cuya expresión no sea variable a lo largo de la diferenciación) el propio algoritmo es incapaz de filtrarlo por lo que su comportamiento es oscilante e incluso aberrante en las primeras etapas. También indicaría que la función objetivo es en exceso laxo, por el contrario una rápida selección y una gran oscilación en pocas generaciones (manteniéndose por encima de un hipotético 'mínimo local') es indicativo de una función objetivo más restrictiva.

4.3.1. El algoritmo genético consigue la convergencia al cabo de 200 generaciones

Para el establecimiento del prototipo del algoritmo genético se ha probado la influencia en la convergencia de los parámetros siguientes: longitud de los cromosomas-virtuales L_{vc} (Es decir, el número de genes-virtuales por cromosoma-virtual), número cromosomas-virtuales N_{vc} y número de generaciones N_{gr}.

En la evolución de las distintas combinaciones de número de genes-virtuales (**Fig. 4.4**) observa-

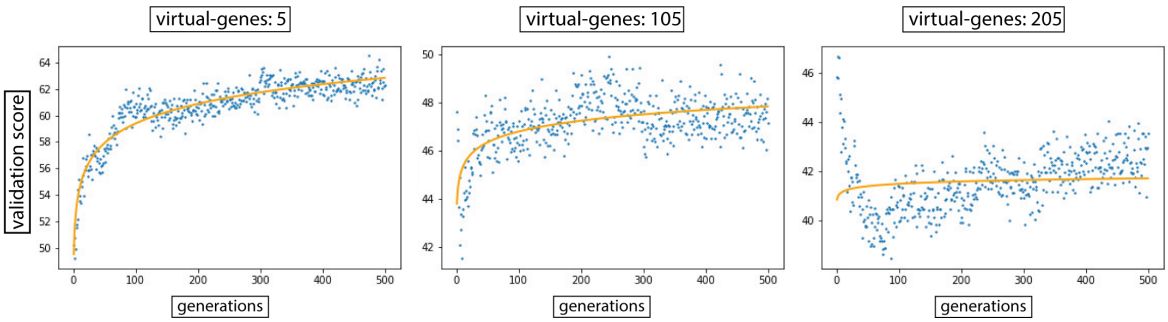


Figura 4.4: **Curvas de convergencia del algoritmo genético utilizando el método de validación de DB. Comparativa de número de genes-virtuales por número cromosoma-virtual.** Se muestran de izquierda a derecha el valor mínimo de genes-virtuales (5), el valor medio (105) y el valor alto (205). En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos graficados.

mos que el algoritmo genético consigue una convergencia para valores bajos y medios de números de genes-virtuales, ya que podemos observar una evolución de la media de la puntuación con forma de curva exponencial positiva convexa. En los valores medios se observa una mayor oscilación una vez alcanzada la convergencia, aún así se puede apreciar una curva exponencial positiva convexa. Sin embargo, valores muy altos del número de genes-virtuales por cromosoma-virtual genera una curva exponencial negativa cóncava y se estabiliza generando una oscilación. Esto nos indica que la introducción de un número excesivo de genes-virtuales por cromosoma-virtual genera una amplificación del ruido superando la capacidad de filtrado del algoritmo genético debido en parte a una laxitud alta de la función de DB.

El comportamiento observado al variar el número de los cromosomas-virtuales (**Fig. 4.5**), se puede observar que en todos los casos el algoritmo converge. Se observa que al aumentar el número de cromosomas-virtuales el algoritmo genético muestra un comportamiento oscilante en las primeras generaciones y al aumentar el número de cromosomas-virtuales necesita un mayor número de generaciones para alcanzar un punto estable en el que comienza a obtener resultados más altos en el valor de la función objetivo.

El aumento de la cantidad de genes que intervienen en el algoritmo genético ya sea mediante repeticiones como genes-virtuales dentro de los cromosomas-virtuales como por el número de cromosomas-virtuales del algoritmo, provoca una situación que fuerza al algoritmo a aumentar el número de generaciones para converger, eliminando así los genes que introducen no contribuyen a la optimización de la función objetivo. Por ello es importante intentar utilizar el número mínimo posible de genes-virtuales respetando la probabilidad de que al menos 1 copia de cada gen participe en el algoritmo genético.

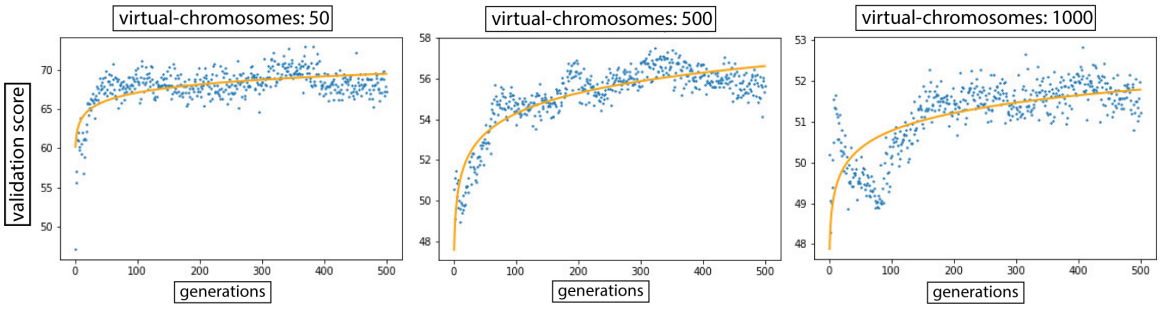


Figura 4.5: **Curvas de convergencia del algoritmo genético utilizando el método de validación de DB. Comparativa de número total de cromosomas-virtuales.** Se muestran de izquierda a derecha el valor mínimo de cromosomas-virtuales (50), el valor medio (500) y el valor alto (1000). En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos graficados.

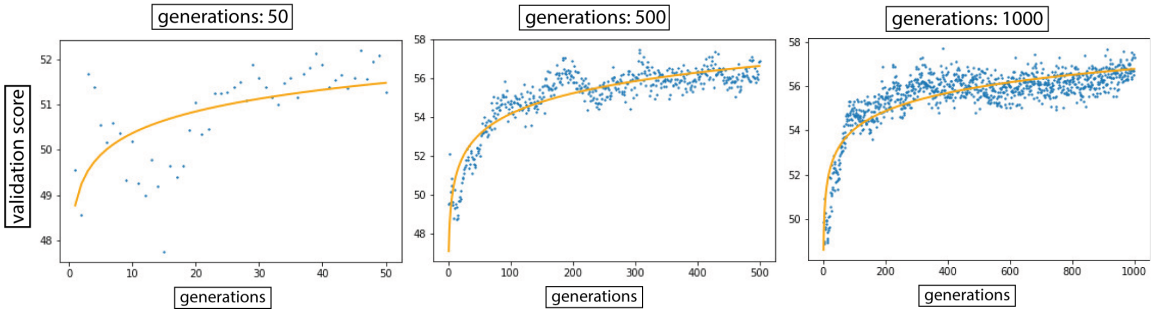


Figura 4.6: **Curvas de convergencia del algoritmo genético utilizando el método de validación de DB. Comparativa del número de generaciones.** Se muestran de izquierda a derecha el comportamiento de la media de las puntuaciones para una combinación de 5 genes-virtuales por 500 cromosomas-virtuales, a lo largo de las generaciones, paneles de convergencia a las 50, 500 y 1000 generaciones. En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos graficados.

Por otro lado, observamos que un mayor número de generaciones (**Fig. 4.6**) provoca una convergencia del algoritmo genético y esta se mantiene con pequeñas oscilaciones. Para el prototipo se observó el alcance de la convergencia a lo largo de las 200 generaciones, a partir de las cuales la puntuación media obtenida por los cromosomas-virtuales seleccionados oscila entre los valores de decrecimiento de 54 %- 57 %.

4.3.2. El método de DC y de DEC alcanzan un punto de inflexión en torno a las 50 generaciones

Ambos métodos basados en creodos consiguen la convergencia del algoritmo genético. Sin embargo se observan diferencias en la forma de convergencia y cómo responden a los datos de

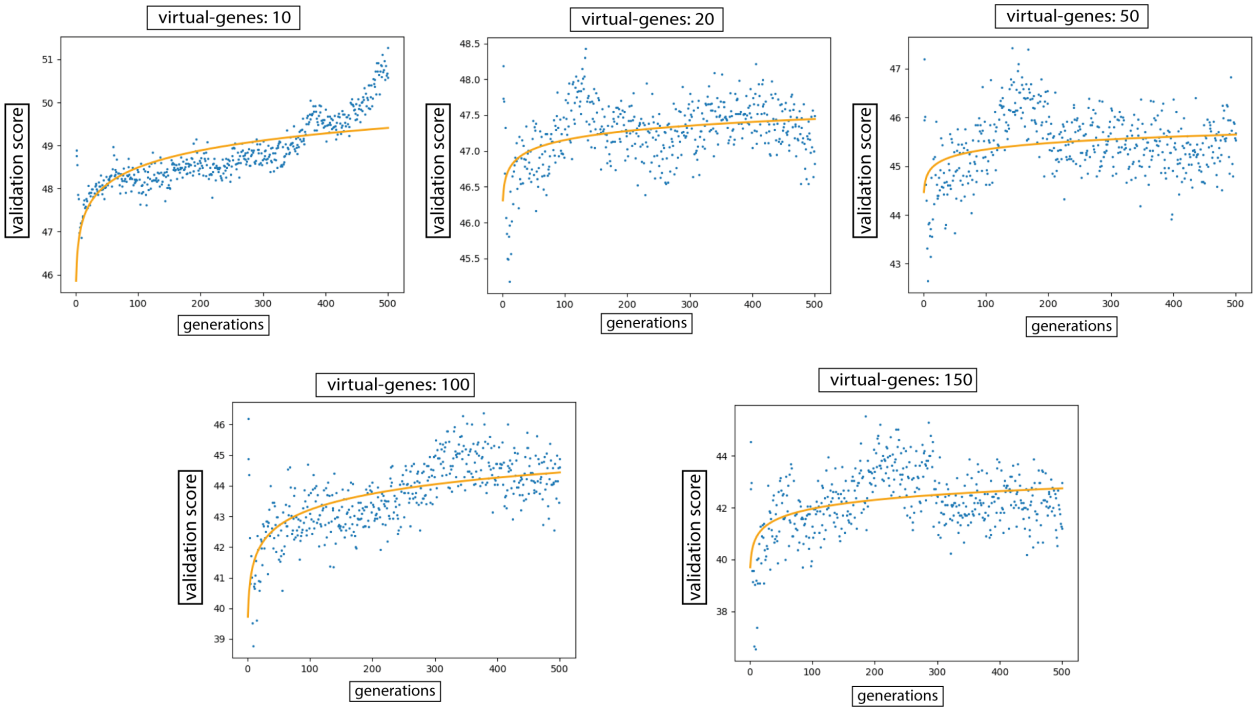


Figura 4.7: **Curvas de convergencia de las distintas pruebas realizadas para el método por DC. Para la totalidad de los genes** Los paneles muestran los distintos valores de genes-virtuales utilizados (combinación de cromosomas-virtuales en la **Tabla 4.1** para un total de 23520 genes de entrada. En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos representados.

entrada.

Al representar las curvas de convergencia del método por DC se observan curvas decrecientes al introducir aquellos genes seleccionados que tienen una mayor variabilidad de expresión a lo largo de las muestras (**Fig. 4.7**). Es por ello que se realizó un análisis adicional con todas las muestras, donde se observa una convergencia creciente y una estabilización en valores medios mayores (**Fig. 4.8**). Esto nos indica que para este método la eliminación del 60 % de los genes que más varían supone una pérdida de información relevante en le caso de utilizar la función objetivo DC.

En el caso de la función objetivo DEC, cada uno de los creodos se muestra por separado. Esto nos indica que para la mayoría de los creodos este método consigue una curva creciente al representar las curvas de convergencia (**Fig. 4.9**). Es reseñable que para el creodo que termina en la diferenciación de macrófagos este método provoca los resultados más aberrantes. Observamos que aunque durante las primeras generaciones parta de una puntuación de validación de 0 (o cercana a 0) finalmente consigue generar combinaciones de genes-virtuales que obtienen puntuaciones reseñables. Esto nos indica que está sucediendo una selección, pero observamos

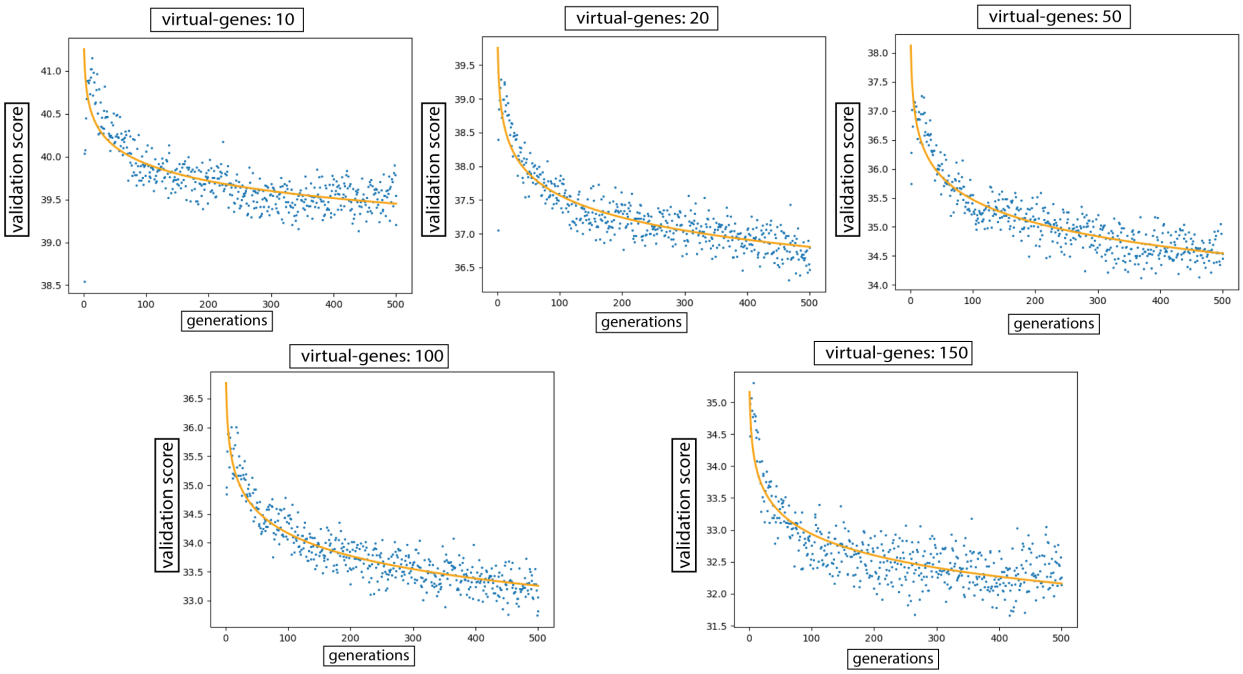


Figura 4.8: **Curvas de convergencia de las distintas pruebas realizadas para el método por DC. para los genes con mayor variabilidad a lo largo de las muestras.** Los paneles muestran los distintos valores de genes-virtuales utilizados (combinación de cromosomas-virtuales en la **Tabla 4.1**) para un total de 9408 genes de entrada. En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos representados.

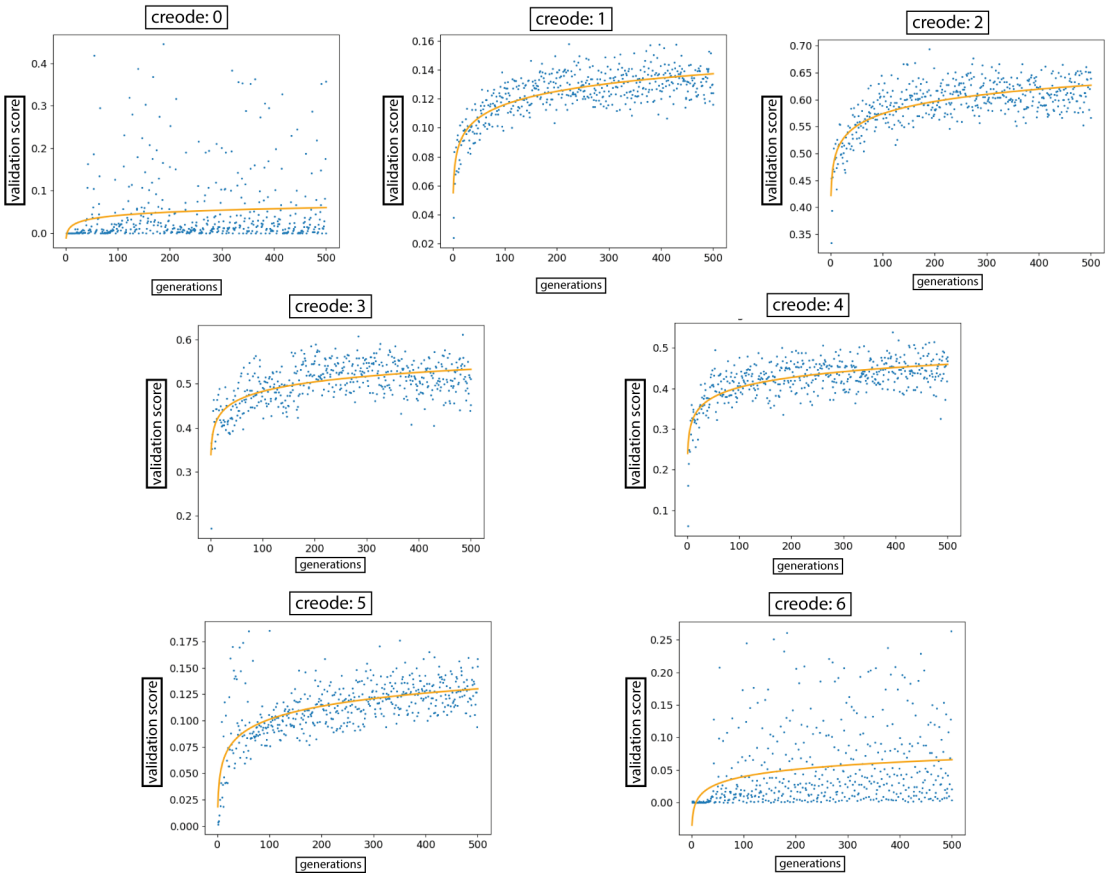


Figura 4.9: **Curvas de convergencia de la prueba de 100 genes-virtuales realizada para el método DEC.** Los paneles muestran los distintos creodos para una combinación de 100-genes dato y 100 cromosomas-virtuales para un total de 9408 genes de entrada. En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados cuyos genes-virtuales conformarán el *pool* de datos de la siguiente generación. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos graficados.

que una vez se ha alcanzado un mínimo la oscilación de la media de los resultados del algoritmo genético a lo largo de las generaciones es mayor (**Fig. 4.9**).

Los distintos creodos se muestran a continuación en la **Tabla 4.3**.

Codificación	Tipo celular atractor del creodo (con sinónimos)
0	macrophage dendritic cell progenitors, mdp, common myeloid progenitor cells, cmp, ...
1	neutrophils
2	mature b cells, immature b cells
3	t cytotoxic cells
4	t helper cells
5	monocytes
6	reticulocytes, erythroblasts, normoblast

Tabla. 4.3: **Listado de creodos con la codificación utilizada en la Fig. 4.9.** Se muestra el tipo celular terminal para cada caso junto con sus sinónimos según la ontología generada en FOntCell

A la luz de estos resultados, se observan diferencias en el comportamiento de ambas funciones objetivo. Para la función objetivo DEC es importante realizar una reducción de datos de entrada ya que es muy exigente a nivel computacional. Durante las primeras generaciones este método elimina los genes-virtuales que otorgan peores puntuaciones para la función objetivo y en torno a la generación $\backsim 50$ el método alcanza un punto de inflexión. En unos genera una asíntota y en otros sucede una oscilación, indicándonos que el algoritmo genético ya no es capaz de optimizar en mayor medida la función objetivo.

En el otro caso, el DC, observamos un efecto similar al introducir la totalidad de los datos. En torno a la generación 50 se observa también el punto de inflexión que da lugar a una oscilación de los resultados fruto de que la función objetivo ya no puede seguir optimizando más el conjunto de genes-virtuales que quedan en el algoritmo.

En cuanto a la combinación de parámetros observamos que valores pequeños de longitud de cromosomas-virtuales (L_{vc}) y mayor número de cromosomas-virtuales (N_{vc}) generan menor oscilación una vez el algoritmo genético consigue la convergencia.

4.3.3. El análisis de GO muestra que la función objetivo DEC selecciona genes más concordantes con el tipo de muestra seleccionada

El análisis GO mostró una diferencia clara en los resultados en función del método utilizado. Por un lado, la función objetivo DC sólo consiguió datos estadísticamente relevantes para aquellas combinaciones en las que se ha realizado una selección previa de los genes más variables y concretamente se encuentran datos estadísticamente significativos cuando utilizamos la combinación de $L_{vc} = 100$ genes-virtuales por cromosoma-virtual . Por otro lado, la función objetivo DEC consigue resultados estadísticamente significativos con las combinaciones de $L_{vc} = 20; L_{vc} = 50; L_{vc} = 100$ genes-virtuales por cromosoma.

La combinación de $L_{vc} = 20$ genes-virtuales parecen mostrar que son muy pocos genes por

Función biológica enriquecida en términos GO para DEC: genes-virtuales = 20	*p*-valor
B cell receptor signaling pathway	2.77E-07

Tabla. 4.5: **Resultados estadísticamente significativos del enriquecimiento en términos por ontología de genes.** Para la prueba del método de función de validación de decaimiento exponencial creódico con la combinación de 20 genes-virtuales y 450 cromosomas-virtuales, partiendo de un total de 9408 genes.

Función biológica enriquecida en términos GO para DEC: genes-virtuales = 50	*p*-valor
peptide antigen assembly with MHC class II protein complex	2.06E-05
peptide antigen assembly with MHC protein complex	2.12E-07
MHC protein complex assembly	2.12E-07
MHC class II protein complex assembly	2.06E-05
immunoglobulin production involved in immunoglobulin-mediated immune response	2.51E-05
B cell receptor signaling pathway	9.54E-06
antigen receptor-mediated signaling pathway	7.27E-07
immune response-activating cell surface receptor signaling pathway	1.32E-05
immune response-regulating cell surface receptor signaling pathway	1.42E-05
immune response-regulating signaling pathway	1.72E-05
immune response-activating signaling pathway	1.70E-05

Tabla. 4.6: **Resultados estadísticamente significativos del enriquecimiento en términos por ontología de genes.** Para la prueba del método de función de validación de decaimiento exponencial creódico con la combinación de 50 genes-virtuales y 200 cromosomas-virtuales, partiendo de un total de 9408 genes.

cromosoma-virtual como para identificar un resultado estadísticamente significativo y que se ajuste a las hipótesis de los distintos métodos de la función objetivo. El otro extremo, $L_{vc} = 150$ genes-virtuales, observamos que en ninguna de las opciones se observan resultados estadísticamente significativos, y genera también problemas de convergencia en la función objetivo DEC al ser muy restrictivo para algunos de los creodos. En cuanto al análisis del proceso biológico estadísticamente significativo, aislado de cada una de las combinaciones de genes-virtuales y de los métodos, observamos los resultados mostrados en las **Tablas 4.4, 4.5, 4.6 y 4.7**.

Función biológica enriquecida en términos GO para DC: genes-virtuales = 100	*p*-valor
translation	2.10E-06
organic substance biosynthetic process	6.49E-06
biosynthetic process	8.29E-06
peptide biosynthetic process	4.16E-06
cellular nitrogen compound biosynthetic process	1.74E-05
cellular biosynthetic process	7.69E-06
organonitrogen compound biosynthetic process	8.51E-07

Tabla. 4.4: **Resultados estadísticamente significativos del enriquecimiento en términos por ontología de genes.** Para la prueba del método de función de validación de decrecimiento creódico con la combinación de 100 genes-dato y 100 cromosomas-dato, partiendo de un total de 9408 genes.

En general, se observa que valores intermedios-altos de número de genes-virtuales satisfacen

Función biológica enriquecida en términos GO para DEC: genes-virtuales = 100	***p*-valor**
phosphatidylinositol biosynthetic process	8.12E-05
glycerophospholipid biosynthetic process	6.45E-05
glycerolipid metabolic process	4.99E-05
primary metabolic process	9.74E-10
metabolic process	7.53E-10
organic substance metabolic process	2.99E-10
cellular metabolic process	1.85E-07
cellular biosynthetic process	2.70E-05
biosynthetic process	5.65E-06
organic substance biosynthetic process	3.30E-06
organic substance catabolic process	3.48E-06
catabolic process	1.22E-06
positive regulation of response to stimulus	4.60E-05
regulation of response to stimulus	7.42E-05
organonitrogen compound metabolic process	9.51E-06
nitrogen compound metabolic process	1.94E-07
macromolecule metabolic process	1.62E-06
UNCLASSIFIED	1.57E-05
G protein-coupled receptor signaling pathway	1.78E-05
detection of chemical stimulus involved in sensory perception of smell	1.32E-05
detection of chemical stimulus involved in sensory perception	4.06E-06
detection of stimulus involved in sensory perception	7.32E-05
sensory perception	5.54E-05
nervous system process	2.12E-07
system process	2.03E-06
detection of chemical stimulus	3.24E-05
sensory perception of chemical stimulus	1.46E-05
sensory perception of smell	9.62E-06

Tabla. 4.7: **Resultados estadísticamente significativos del enriquecimiento en términos por ontología de genes.** Para la prueba del método de función de validación de decaimiento creódico con la combinación de 100 genes-virtuales y 100 cromosomas-virtuales, partiendo de un total de 9408 genes.

mejor la función objetivo y generan un mayor número de resultados estadísticamente significativos. Así como el método por DEC consigue más resultados estadísticamente significativos y con sentido biológico que los generados con el método de DC.

4.4. Discusión

Con las muestras normalizadas en su conjunto y sus relaciones ontológicas hemos podido diseñar un algoritmo genético que nos permite optimizar una función objetivo para la obtención de un conjunto de genes que generen un paisaje de Waddington.

Podemos constatar que, aunque el método no exige una preselección de los genes de entrada, esta inicialización mejora considerablemente el rendimiento y los resultados. Por ello se realiza una preselección de aquellos genes cuya expresión es más variable. Esto también abre la puerta de poder elegir conjuntos de genes escogidos por el usuario: si por ejemplo, quisiéramos observar qué genes dentro de una familia de genes son aquellos que en conjunto mejor construyen un paisaje de Waddington.

En cuanto a los parámetros del algoritmo genético, se ha observado su funcionamiento con diferentes combinaciones de número genes-virtuales por cromosoma-virtual (L_{vc}) y número de cromosomas-virtuales (N_{vc}). En ambos casos, un número alto de cromosomas-virtuales (y bajo de genes-virtuales por cada cromosoma-virtual) o un número alto de genes-virtuales por cromosoma-virtual (y bajo de cromosomas-virtuales) nos supone una mayor carga computacional y generalmente no proporciona resultados estadísticamente significativos en el posterior análisis de GO. En el caso de un número bajo de L_{vc} facilita la optimización de la función objetivo, haciendo que un mayor número de combinaciones sean satisfactorias. También, dificulta distinguir la información relevante de la irrelevante y generando resultados donde no se encuentran términos ontológicos estadísticamente significativos en un análisis de GO. En el caso de una mayor L_{vc} nos encontramos con el caso opuesto: un mayor número de genes-virtuales por cromosoma-virtual dificulta que satisfagan la función objetivo y dado que implica un número más bajo de cromosomas-virtuales, debido a la combinatoria de los genes de entrada que introducimos, podemos encontrarnos con la situación que a lo largo de las generaciones la puntuación de los cromosomas-virtuales pueda ser 0.

Tal y como ha podido observarse, el número de generaciones requeridas en el algoritmo genético son $N_{gr} = 200$ para que la función objetivo del algoritmo genético se estabilice en su asíntota de convergencia. Aunque se mantendrá en torno a las 500 generaciones para presentar un margen de seguridad. Se ha observado en algunas pruebas de convergencia durante las primeras generaciones que la media de la puntuación de los cromosomas-virtuales resultantes pueden oscilar,

incluso describir una curva decreciente, se especula que este fenómeno pueda deberse a una deriva generada por las propias oscilaciones del algoritmo genético al introducir un conjunto de datos que genere resultados similares en el algoritmo genético y también deberse a un elevado número de genes que no son útiles para la optimización de la función objetivo. Puede que los datos de transcripción del genoma tengan este comportamiento sobre todo para las funciones objetivo por decrecimiento DB y DC. Para la función objetivo DEC se observa también un hecho similar en algunos creodos, en los cuales durante las primeras generaciones se obtienen valores de putnuación media de 0 (o próximos a 0) y cuando el algoritmo se deshace de los genes que no maximizan la función objetivo comienza a obtener resultados mejores y a estabilizarse en una oscilación.

La función objetivo del algoritmo genético selecciona el conjunto de genes que tengan un comportamiento decreciente a lo largo del paisaje de Waddington y con un gradiente en función de la diferenciación celular. En el caso de la función objetivo DB nos encontramos con un método cualitativo binario que mide las relaciones que decrecen o no decrecen. Adicionalmente, la función objetivo DB va a seleccionar un mayor número de relaciones celulares decrecientes, entendiendo por decreciente que el status transcriptómico (St) del conjunto de genes a evaluar sea mayor en el ascendente que en el descendente. En el caso de la función objetivo DC se incorpora información cuantitativa, ya que aquellas relaciones que tengan un mayor decrecimiento en cuanto a salto del status transcriptómico (St) de los ascendentes al descendiente se seleccionarán por el algoritmo genético ya que optimizan la función objetivo. La función objetivo DC no sólo selecciona positivamente el decrecimiento, sino que además selecciona aquellos genes que tengan un mayor grado de decrecimiento (una mayor pendiente), y penalizará aquellos cambios en la tendencia decreciente dentro de los creodos, sobre todo en aquellos que el cambio sea abrupto.

El principal problema que hallamos en las funciones objetivo DB y DC a la hora de generar un paisaje de Waddington es que permiten que se seleccionen algunas relaciones crecientes, y aunque DC es más robusta a la hora de evitar la selección de relaciones crecientes, es cierto que es más permisivo a este fenómeno que la función objetivo DEC. Por lo tanto, las funciones objetivo DB y DC podrían seleccionar conjuntos de genes que incluyan relaciones crecientes lo que se traduce en una deformación del resultado del paisaje de Waddington.

Por contrapartida, la función objetivo DEC es mucho más restrictiva ya que se evalúa el ajuste a una curva exponencial y se asigna una puntuación de 0 a aquellas pendientes positivas. Esto provoca que haya un número relativamente alto de cromosomas-virtuales que tengan una puntuación de 0 lo que provoca la presencia de algunos cromosomas-virtuales con baja puntuación no sean eliminados con el paso de las generaciones. También, sería mucho más restrictivo si se tuvieran en cuenta todos los creodos o los ascendentes más cercanos (como en la función objetivo DC), por lo que se decidió dividir el conjunto de relaciones por creodos y evaluarlas por

separado, aunque esto implique una menor optimización del algoritmo genético (en términos de carga computacional y tiempo de ejecución), ya que debe evaluar en paralelo cada uno de los creodos. Aunque computacionalmente requiere de una mayor carga, la evaluación de los creodos por separado con el algoritmo genético puede permitir en un futuro un estudio más minucioso de los resultados.

Adicionalmente, se observa que la función DEC genera resultados que tienen un mayor significado biológico. Es por ello, que se utilizará este método en siguientes implementaciones del algoritmo genético, descritas en el capitulo siguiente.

Recapitulando, hemos desarrollado e implementado un algoritmo genético que puede maximizar una función objetivo. Hemos observado que las mejores combinaciones para el conjunto de genes que introducimos en el algoritmo, son un número medio-alto de genes-virtuales por cromosoma-virtual siempre y cuando no se traduzca en un número muy bajo de cromosomas-virtuales. Y también, hemos observado que la función objetivo DEC mejora los resultados de convergencia y son biológicamente más relevantes que los otros métodos.

Capítulo 5

paisaje de Waddington transcriptómico integral

5.1. Introducción

En este punto, nos encontramos con los elementos necesarios para la construcción de un paisaje de Waddington integral. El paisaje de Waddington que vamos a construir en este capítulo va a utilizar como referencia la ontología generada por FOntCell (Cabau-Laporta et al., 2021), el cual es el resultado de la fusión de las ontologías de CELDA y LifeMap. Esta ontología es el consenso de ambas ontologías y contiene los nombres estandarizados de los tipos celulares y sus relaciones de desarrollo.

Los datos recolectados para la construcción del paisaje de Waddington corresponden a datos transcriptómicos de arrays de *Affymetrix* de humano. En la actualidad, con la explosión de las técnicas de *single-cell* (Grün, Oudenaarden, 2015);(Papalexi, Satija, 2017);(Chai, 2022) encontraríamos datos más precisos a la hora de discernir diferentes tipos celulares en un mismo tejido (Chai, 2022), aunque continuaríamos encontrándonos con el problema de la falta de estándares entre tipos celulares, debido, principalmente a que el etiquetado de los clusters de *single-cell* debe realizarse de forma manual. Aunque se espera que el avance en el proyecto *Human Cell Atlas* consiga una mejor categorización de los tipos celulares y los estadios intermedios entre ellos (Chai, 2022), donde la creación del paisaje de Waddington podría contribuir a esa tarea. Como la idea original del proyecto del paisaje de Waddington transcriptómico contemplaba el uso de tipos celulares humanos de la plataforma GPL570 de *Affymetrix* se decidió continuar con la idea original ya que todo el software que se ha desarrollado iba en esa línea. Aun así, obtener un paisaje de Waddington con este tipo de datos puede ser la primera piedra en un proyecto posterior de cons-

trucción de un paisaje de Waddington con datos *single-cell* y serviría como un primer prototipo válido.

Tal y como se comentó en capítulos anteriores, la construcción del paisaje de Waddington transcriptómico consiste en seleccionar un conjunto de genes que puedan construir y representar la estructura del paisaje de Waddington. Para ello, se ha diseñado un algoritmo genético que pueda seleccionar un conjunto de genes a partir de la información transcriptómica y jerárquica de las muestras que introduzcamos. Este algoritmo genético ha sido puesto a prueba con un conjunto de datos compuesto por tres GSEs diferentes todos ellos centrados en datos hematopoyéticos, lo cual nos ha permitido observar y optimizar el funcionamiento del algoritmo genético, como interaccionan los parámetros entre sí, como interacciona e influye en todo el proceso las diferentes funciones de aptitud y como optimizar sus parámetros.

Los tipos celulares hematopoyéticos nos presentan un conjunto de datos jerárquicamente bien definidos (Dzierzak, Bigas, 2018) y utilizados en algunos trabajos centrados en los procesos de diferenciación y desarrollo de los mismos (Mabbott et al., 2013);(Rapin et al., 2014). Con lo cual, disponemos de un conjunto inicial de datos mejor definidos con los que evaluar el funcionamiento del algoritmo genético en una prueba de concepto. En la prueba de concepto nos hemos centrado en los aspectos técnicos del proceso, pero atendiendo también al contenido biológico de los resultados. En esta línea, hemos evaluado cuál de los diferentes métodos generan información más relevante a nivel biológico en su conjunto, no solo por la cantidad de términos ontológicos significativamente relevantes (ver Tablas4.4, 4.7, 4.6 y 4.5) sino por aquellos términos relacionados con los tipos celulares hematopoyéticos.

En el presente capítulo, disponemos de un conjunto de datos basados en la totalidad de los tipos celulares categorizados en la ontología resultante de la fusión de CELDA y LifeMap mediante la herramienta FOntCell. Estos datos han sido recolectados con el objetivo de cubrir el máximo posible de tipos celulares disponibles.

En contraste con el anterior capitulo, se utiliza un total de muestras de un orden de magnitud mayor que en la prueba de concepto y también en el número de creodos. Esto supone un reto a la hora de afrontar los resultados y escalar todos los procesos que intervienen en la selección de aquellos genes que construyan el paisaje de Waddington.

Los resultados obtenidos del algoritmo genético generan una lista de genes para cada uno de los creodos. Para evaluar la calidad de dichos conjuntos de genes, se realizarán dos análisis, por un lado, el análisis de GO (Mi et al., 2012) tal y como se ha realizado en la prueba de concepto, pero atendiendo en esta ocasión a posible información biológica relevante en los términos los cuales nuestra muestra está enriquecida para cada creodo. También, se complementará con un análisis

utilizando datos de tipos celulares públicos en la base de datos de GSEA (*Gene Set Enrichment Analysis*). GSEA es una herramienta cuyo objetivo consiste en buscar asociaciones de un set de genes S con una clase fenotípica teniendo listas de genes L asociados a cada fenotipo (Mootha et al., 2003);(Subramanian et al., 2005), GSEA pone a disposición pública una serie de fenotipos y estados celulares asociados a listas L de genes. Para nuestro análisis utilizaremos la lista C8 de GSEA que está basada en tipos celulares analizados con *single-cell*. En esta base de datos cada tipo celular presenta una lista de genes como su firma genética de aquellos genes asociados a un determinado tipo celular. La base de datos contiene un total de 830 firmas distintas, asociadas a las muestras de distintos tipos celulares de distintos trabajos, donde nos encontramos duplicaciones de tipos celulares pero pertenecientes a distintos trabajos y/o tejidos.

Actualmente, la base de datos de GSEA en lo referente a tipos celulares y su firma genética está construida mediante datos de *single-cell*. Aunque no es el tipo de dato con el que estamos trabajando confiamos en que las firmas genéticas en el caso de *single-cell* y *bulk-cell* deberían ser similares al referirnos a un tipo celular concreto.

5.2. Materiales y Métodos

5.2.1. Muestra y relaciones ontológicas

Para la construcción del paisaje de Waddington se utilizó la jerarquía de diferenciación celular establecida tras la fusión de CELDA y LifeMap utilizando FOntCell. Después se descargaron los datos de forma automática del repositorio de GEO para la plataforma GPL570, siguiendo la metodología descrita en el **capítulo 3**.

La extracción de las muestras para la construcción del paisaje de Waddington se realizó obteniendo el máximo posible de tipos celulares disponibles en la base de datos de GEO de la plataforma GPL570, seleccionando todos los datos que consiguieron detectar ambos algoritmos de forma secuencial. A continuación, se realizó la normalización conjunta de los datos y se procedió con aplicar el software de *ComBat* para la reducción del *batch-effect* (Johnson, Li, 2007).

Dada el número actual de datos y con vistas a una futura escalada de dicho número, los pasos adicionales han consistido en la construcción y ejecución de algoritmos para identificar de forma automática los distintos creodos y así facilitar en el futuro el análisis de los resultados. También, de cara a facilitar el análisis de los resultados, se ha tomado como referencia el tipo celular terminal de cada uno de los creodos y se ha procedido a identificar el linaje celular al que pertenece (cigoto, mesodermo, endodermo y ectodermo) y lo que hemos denominado como 'compartimento celular' que responde a una serie de tejidos diferenciados que se interrelacionan por el linaje y la diferenciación celular, los cuales se reflejan en el árbol de desarrollo *in vivo* de LifeMap Discovery

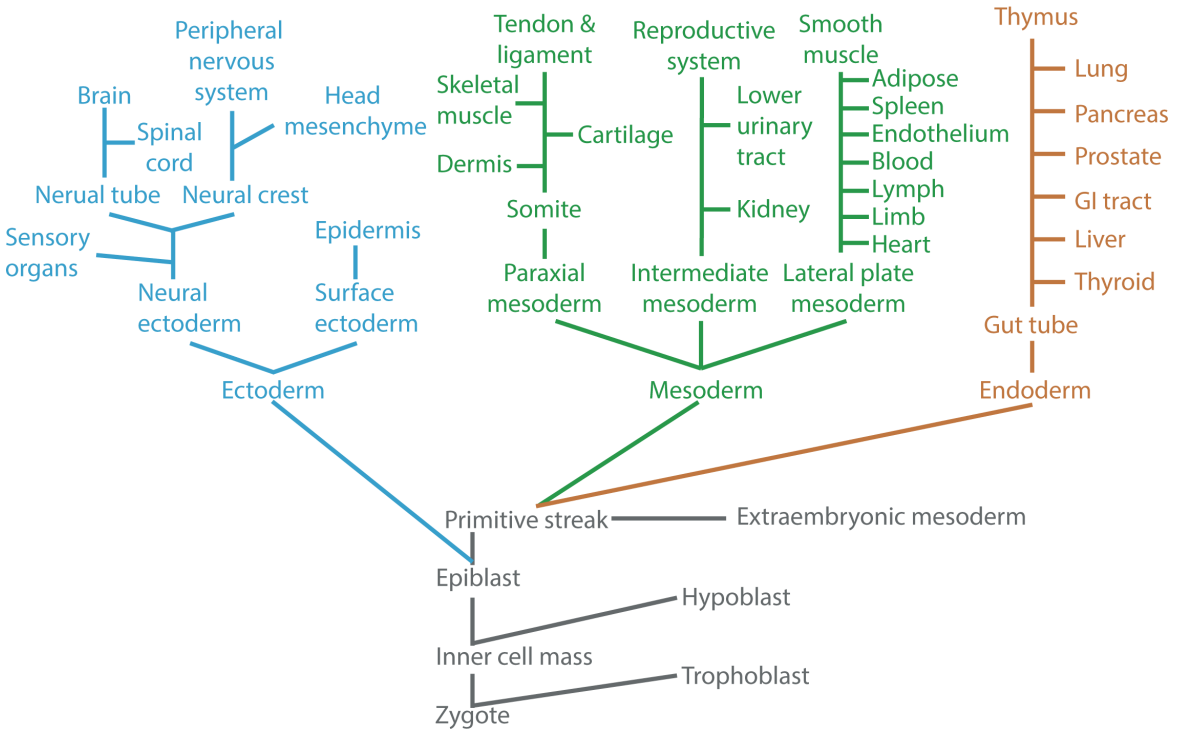


Figura 5.1: **Esquema de desarrollo celular**. Adaptado de la imagen de LifeMap Discovery (Edgar et al., 2013), donde se muestran los 4 linajes, separados por colores (gris para cigoto, azul para ectodermo, verde para mesodermo y naranja para endodermo) y los 45 compartimentos celulares en los que se dividen los diferentes linajes.

(Edgar et al., 2013) de la figura **Fig. 5.1**.

5.2.2. Escalado del algoritmo genético para la construcción del paisaje de Waddington y la optimización de parámetros

Atendiendo a los resultados obtenidos en el Capítulo 3, para la construcción del paisaje de Waddington se optará por utilizar la función objetivo que utiliza el método de maximización del DEC. Aunque es la que mejor resultados obtiene, esta función es la que mayor carga computacional requiere y en el caso de la construcción del paisaje de Waddington, nos encontramos con un total de 73 creodos diferentes, por lo que supone un total de 10 veces más creodos que en el caso de la construcción del paisaje de Waddington hematopoyético utilizado para generar la prueba de concepto. Ello implica un aumento del tiempo de cálculo de cada generación del algoritmo genético en especial cuando se lanzan procesos con un número grande de cromosomas-virtuales. Para mejorar la optimización se implementa una configuración en paralelo y se añade una funcionalidad secuencial para poder observar la evolución de los datos a lo largo de determinadas generaciones.

En cuanto a los parámetros iniciales del algoritmo genético se va a optar por las combinaciones de número de genes-virtuales y número de cromosomas-virtuales basados en las mejores configuraciones obtenidas en la prueba de concepto, ya que el número de genes a evaluar se mantiene, lo que aumenta es el número de creodos y de tipos celulares. Concretamente, se va a utilizar una combinación de parámetros de 50 genes-virtuales con 200 cromosomas-virtuales, 75 genes-virtuales con 150 cromosomas-virtuales y 100 genes-virtuales con 100 cromosomas-virtuales, con el objetivo de utilizar un número de genes-virtuales considerados medio-alto junto con un número de cromosomas-virtuales que proporcione una probabilidad de ocurrencia de mínimo una copia de cada gen para el conjunto de genes-virtuales y cromosomas-virtuales en el algoritmo genético (**Eq. (4.2.12)**). El desglose de los parámetros utilizados para la construcción del paisaje de Waddington integral se muestra en la **Tabla 4.2**.

Tras la ejecución del algoritmo genético observaremos el grado de convergencia para cada uno de los diferentes creodos y a continuación observaremos la proporción de elementos significativos obtenidos para los términos de la ontología de genes de las diferentes combinaciones de genes-virtuales y cromosomas-virtuales para escoger aquella que reúna las mejores características en cuanto a convergencia y número de resultados del análisis de GO.

5.2.3. Estudio de GO

Al igual que en la prueba de concepto y partiendo de las relaciones ontológicas descompuestas en los diferentes creodos, se ha procedido a generar un análisis de GO de acuerdo con el protocolo del repositorio web PANTHER (Mi et al., 2012).

El algoritmo genético genera como resultado una lista de genes del tamaño del número de genes-virtuales para cada uno de los creodos. Esta lista de genes serán los que mejor se ajusten a la función objetivo y por tanto aquellos que mejor construyan el paisaje de Waddington. Por lo tanto, para cada creodo se realizará un análisis de GO y se seleccionarán aquellos que sean estadísticamente significativos.

Para comparar las diferentes combinaciones de conjuntos (50 genes-virtuales, 75 genes-virtuales y 100 genes-virtuales) seleccionaremos el máximo *p*-valor de cada uno de los términos asociados con los resultados de cada creodo tanto aquellos con términos significativos como aquellos que no tengan enriquecimiento de términos de genes significativos. Después se seleccionará para una evaluación más detallada la combinación de genes-virtuales aquella cuyos valores máximos de *p*-valor para cada creodo sean menores en conjunto. Después se analizarán los resultados del análisis de GO en busca de patrones o información de interés biológico.

5.2.4. Análisis comparativo con el registro de tipos celulares de GSEA

Adicionalmente, se ha realizado un análisis adicional del paisaje de Waddington basado en un alineamiento con los tipos celulares existentes en GSEA. El objetivo de este análisis es la de alinear los genes resultantes del algoritmo genético para cada creodo con las huellas genéticas de la base de datos de GSEA. De esta forma buscar coincidencias y seleccionar las mejores coincidencias para cada creodo. Estas coincidencias, aunque no sean exactas con el atractor del creodo, es esperable que tengan similitudes respecto al linaje y/o compartimento celular con el propio atractor.

En el repositorio de GSEA podemos obtener información de un total de 830 tipos celulares y un conjunto de genes que están sobrerrepresentados para ese tipo celular concreto, lo que podemos identificar como una 'huella genética' asociada a cada uno de los tipos celulares.

Por nuestro lado, disponemos de los resultados del algoritmo genético, que son un conjunto de genes que explican la diferenciación para cada creodo obtenidos mediante la función objetivo basada en DEC. Para analizar nuestros datos en función del registro de tipos celulares de GSEA, tomaremos como referencia de cada creodo el tipo celular terminal y de ese tipo celular dispondremos el compartimento celular y el linaje.

Teniendo la lista de genes de cada uno de los creodos generaremos una matriz comparativa entre los distintos tipos celulares de GSEA con los distintos creodos de nuestro paisaje de Waddington. Para cada uno de los creodos se calculará el porcentaje de coincidencia tal y como se muestra en la **Ec. 5.2.4**.

$$coincidencia = \frac{interseccion}{combinacion\ de\ vg} \qquad \text{(Ec. 5.2.1)}$$

Donde la $combinaciondevg$ viene determinado en los parámetros de entrada del algoritmo genético y será el número de genes que tiene como resultado cada uno de los creodos. El término $interseccion$ es un valor que representa el número de genes de un determinado creodo presente en un tipo celular de GSEA independientemente del número de genes que tengan los tipos celulares de GSEA, cuya media se sitúa en 180 genes. Este cociente nos permitirá tener el resultado de $coincidencia$ en tanto por 1.

Una vez generada la matriz de alineamiento los distintos tipos celulares de GSEA con los creodos se seleccionarán para cada creodo los 5 tipos celulares de GSEA cuya puntuación de coincidencia

sea mayor. A continuación, se analizará las coincidencias en cuanto a linaje celular y compartimento del tipo celular final de cada creodo con los tipos celulares de GSEA asignados.

Este método nos permite relacionar los resultados de un creodo con un tipo celular concreto, aunque entendemos que los mecanismos transcriptómicos que correlacionan con la regulación de un creodo no tienen por qué coincidir con los mecanismos transcriptómicos que regulan los tipos celulares terminales hipotetizamos que potencialmente este análisis pueda relacionar creodos con tipos celulares de acuerdo al linaje celular o el compartimento celular final de ese creodo.

5.3. Resultados

5.3.1. El escalado muestra un funcionamiento similar para las tres combinaciones de genes-virtuales, 50, 75 y 100

El algoritmo genético logra alcanzar la convergencia para todos los casos (50 genes-virtuales, 75 genes-virtuales y 100 genes-virtuales), en torno a las 50 generaciones, utilizando los tipos celulares seleccionados para la construcción del paisaje de Waddington y sus relaciones ontológicas junto con la función objetivo basada en el DEC.

En la figura **Fig. 5.2** se observa una tendencia con ajuste logarítmico que refleja la selección con éxito, de los cromosomas-virtuales que mejor se ajustan a la función objetivo, por parte del algoritmo genético. Esta tendencia se observa en las tres combinaciones de genes-virtuales y a lo largo de los distintos creodos, con excepciones en algún creodo (**Fig. 5.3**), en las cuales el comportamiento es en cierta medida más oscilante pero aun así el algoritmo genético consigue una selección correcta y ascendente de la función objetivo en función del número de los cromosomas-virtuales y genes-virtuales con el paso de las generaciones.

No se observan diferencias significativas a la hora de la convergencia entre las combinaciones de genes-virtuales, es por ello que se realizó un análisis de GO. Tal y como se muestra en la figura **Fig. 5.4**, donde se han seleccionado el *p*-valor más pequeño de cada uno de los creodos en cada una de las tres combinaciones de genes-virtuales, se observa una distribución similar para las tres combinaciones de genes-virtuales.

Puesto que a nivel de escalado de datos se observan similares resultados, tanto en convergencia como en cantidad de valores mínimos significativos obtenidos para el análisis de GO, se seleccionará la combinación de 100 genes-virtuales para los análisis siguientes.

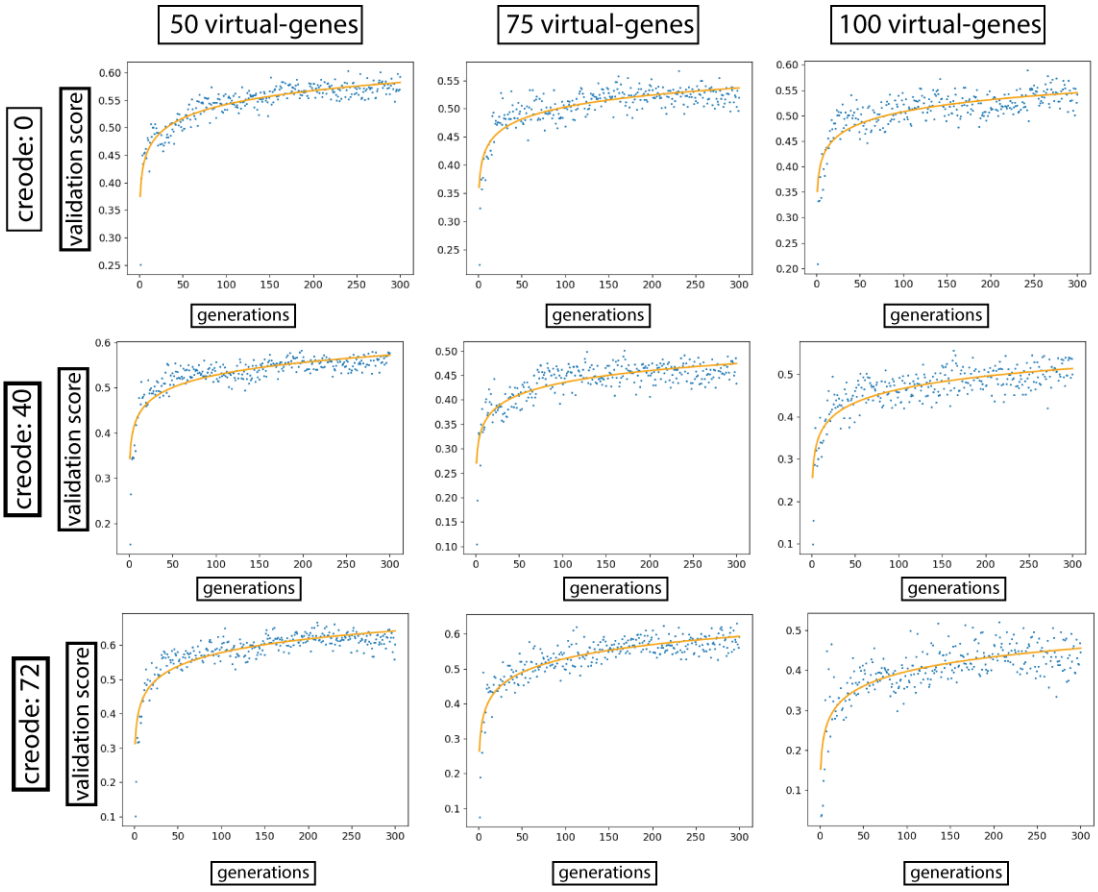


Figura 5.2: **Curvas de convergencia del algoritmo genético para los creodos 0, 40 y 72 con las diferentes combinaciones de genes-virtuales**. El método de la función objetivo es el DEC. Se muestra la puntuación de esos creodos para las diferentes combinaciones de genes-virtuales (50, 75 y 100). En azul se muestra la media de la puntuación de los cromosomas-virtuales seleccionados. La curva naranja es un ajuste mediante una regresión exponencial al conjunto de puntos azules. Se observa que la convergencia ocurre en torno a las 50 generaciones.

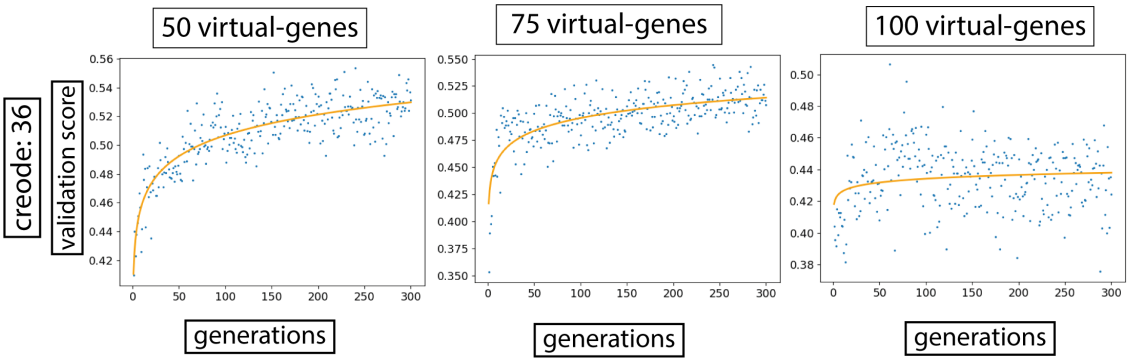


Figura 5.3: **Curvas de convergencia para el creodo 36**. Aunque el funcionamiento es aceptable para la combinación de 50 genes-virtuales, tiene un comportamiento oscilante en los otros casos, sobretodo para la combinación de 100 genes-virtuales. De todas formas se observa cierta convergencia en las combinaciones de genes-virtuales de 75 y 100 en torno a la generación 50, siendo la combinación de 100 genes-virtuales la más oscilante. En azul se muestran las medias de las puntuaciones del conjunto seleccionado de cromosomas-virtuales para cada generación. En naranja se muestra la curva de ajuste mediante una regresión exponencial al conjunto de puntos azules.

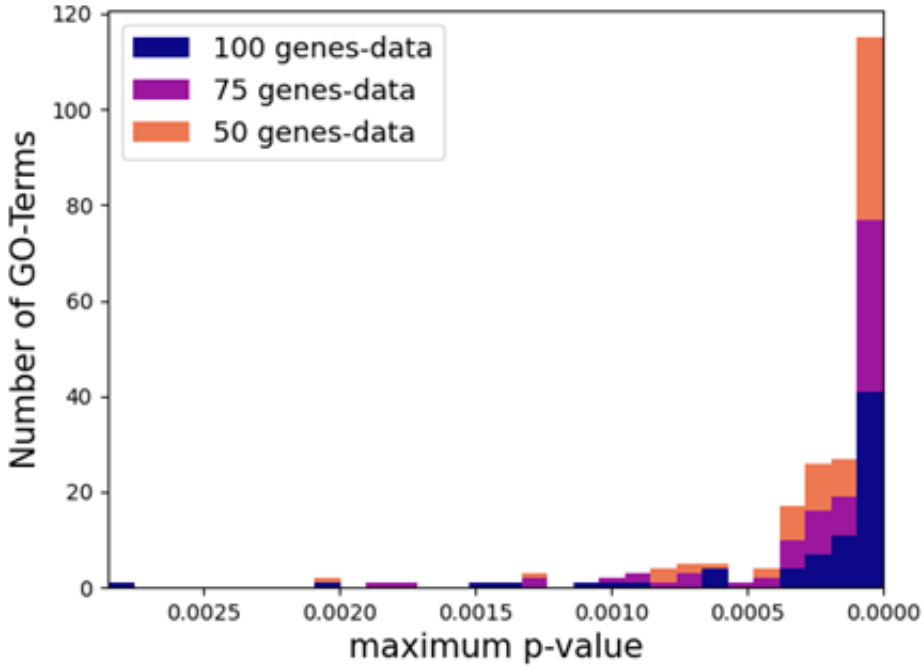


Figura 5.4: **Histograma de la distribución de valores máximos de *p*-valor para cada uno de los creodos** se observa la distribución en función del máximo *p*-valor de los términos del análisis de GO a cada uno de los creodos para cada una de las combinaciones de genes-virtuales: 50 genes-virtuales en naranja, 75 genes-virtuales en magenta y 100 genes-virtuales en azul.

5.3.2. El análisis GO nos muestra un enriquecimiento de genes relacionados con procesos inmunitarios y metabólicos

El resultado del algoritmo genético es una lista de genes para cada creodo, 100 genes por creodo para la combinación que utilizamos, por cada uno de los creodos. De cada una de la lista de genes de cada creodo se realiza un análisis de GO. El resultado de este análisis consiste en la lista completa de términos de la ontología y su *p*-valor para los genes del creodo que se esté estudiando.

Este análisis nos permite relacionar los genes que hemos obtenido con términos y funciones registradas en una ontología de genes, y además conocer el *p*-valor como indicativo de que el enriquecimiento de los genes y términos de la ontología que vamos a analizar es significativo estadísticamente.

Por ello nos vamos a centrar en aquellos términos cuyo *p*-valor es significativo. De los 73 creodos que tiene nuestra ontología, en torno a un 50 % tienen términos de la ontología de genes significativos.

Separando por categorías la función de los términos ontológicos seleccionados podemos observar un enriquecimiento de términos relacionados con procesos inmunitarios y/o típicos del sistema inmune, incluso en creodos que no pertenecen al compartimento hematopoyético o al linaje del mesodermo que es el compartimento del sistema inmune. También encontramos un gran número de procesos metabólicos seleccionados de forma significativa a lo largo de los diferentes creodos. En la **Fig. 5.6** se observa el desglose de los diferentes términos significativos hallados en el análisis de enriquecimiento, categorizados dentro de las distintas categorías y en la **Fig. 5.5** se observa el tamaño de cada una de las categorías, donde resalta especialmente los términos asociados al sistema inmune, seguidos por los términos relacionados con el metabolismo, categoría que engloba a los procesos de catabolismo y anabolismo.

5.3.3. La comparativa con los tipos celulares de GSEA se observa un enriquecimiento de asignaciones de tipos celulares del compartimento hematopoyético

Tras comparar los genes resultantes de cada creodo con los genes asignados a cada tipo celular según la base de datos de GSEA y seleccionar las 5 mejores asignaciones para cada creodo, vemos que se produce un enriquecimiento de asignaciones de tipos celulares relacionados con el compartimento hematopoyético (**Fig. 5.7**) incluso superando la presencia de tipos celulares de otros linajes como ectodermo o endodermo, el linaje con menos asignaciones es el mesodermo si no contamos los tipos celulares de hematopoyesis, es necesario mencionar que no ha habido ninguna asignación para el linaje cigoto (linaje previo a la diferenciación en los tres linajes).

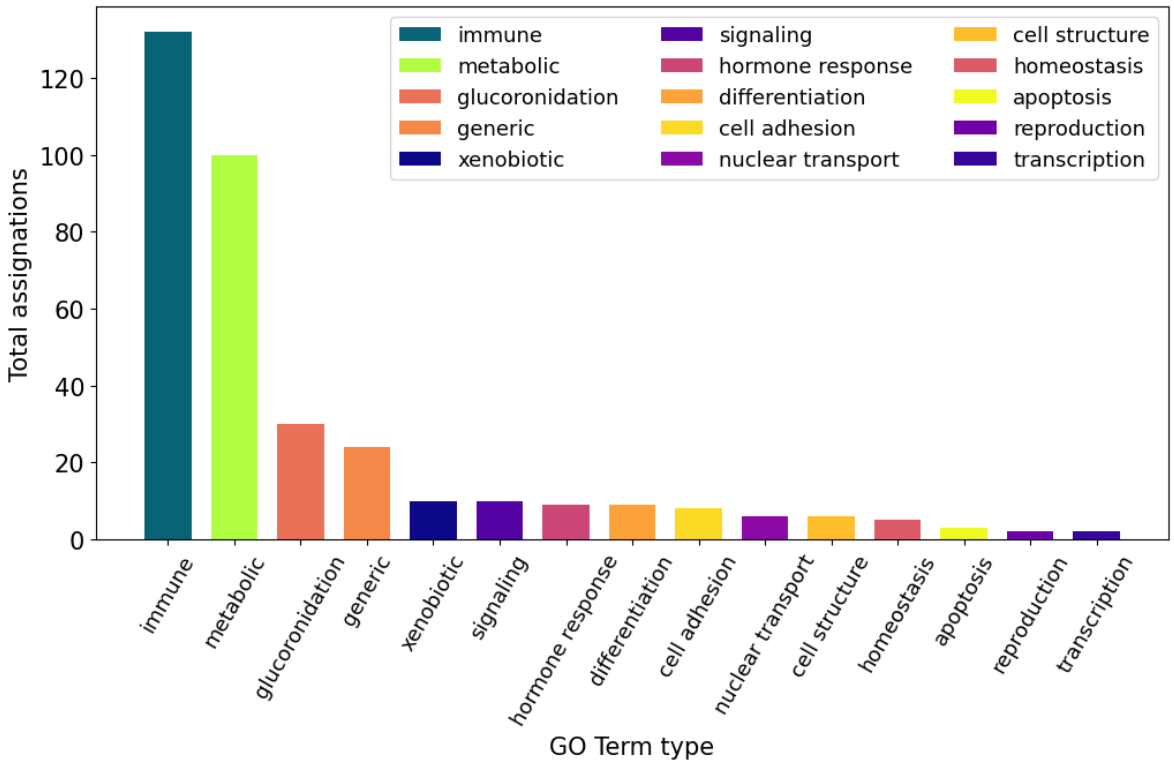


Figura 5.5: **Histograma del total de términos significativos asignados mediante un análisis de GO.** Se observa que las funciones asignadas como 'inmunes' y 'metabólicas' destacaban por encima de las demás en el número de veces que han sido asignadas a los diferentes creodos.

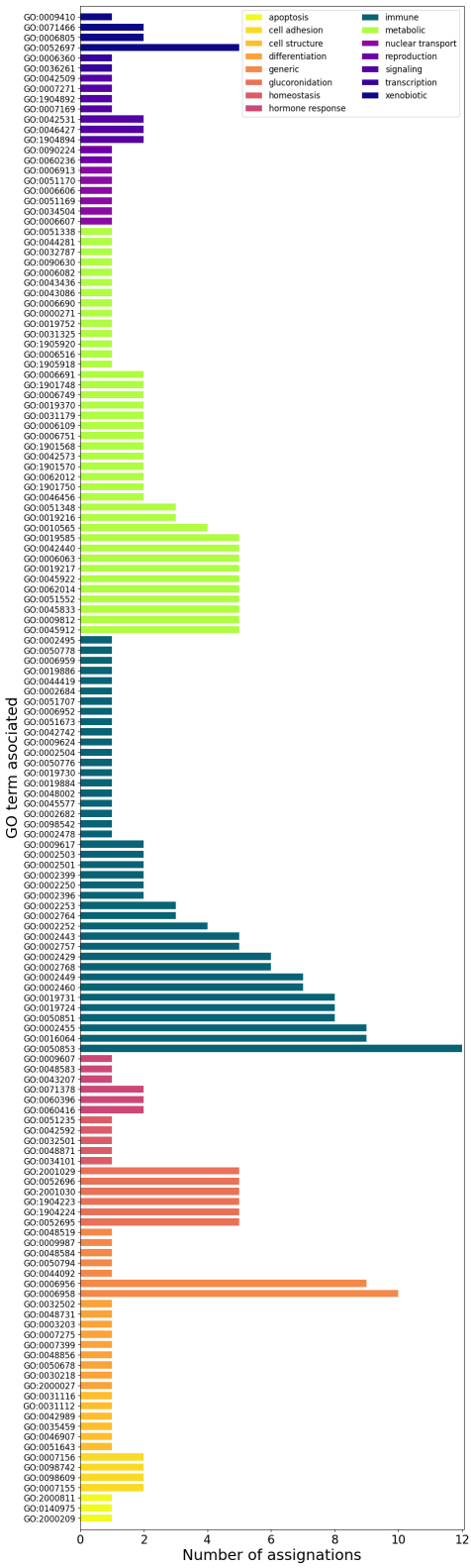


Figura 5.6: **Histograma de la distribución de los diferentes términos significativos hallados en la ontología de genes para los distintos creodos**. El color marca la función celular asignada.

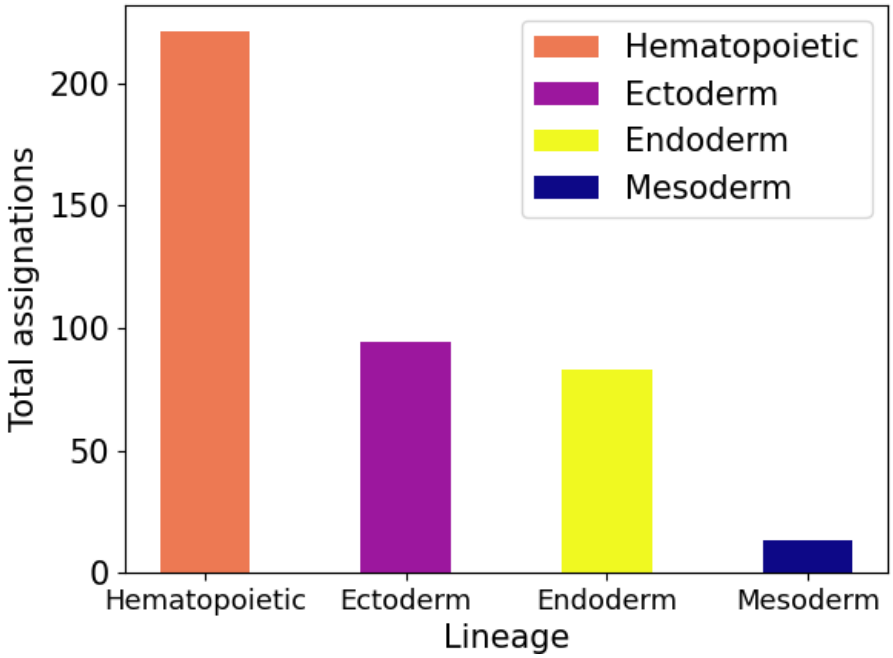


Figura 5.7: **Distribución por linaje de los tipos celulares resultantes del emparejamiento mediante GSEA**. Se destaca los tipos celulares etiquetados como *Hematopoietic* que corresponden al linaje hematopoyético, los cuales se han separado del conjunto del linaje mesodermo para observar mejor su contribución.

Las distintas asignaciones de tipos celulares a nuestros creodos suponen un total de 86 tipos celulares diferentes repartidos a lo largo de los creodos, en la **Fig. 5.8** se observan el número de veces que esos tipos celulares han sido asignados a un creodo (teniendo en cuenta que no puede haber repetición, es decir un mismo tipo celular no puede asignarse dos veces a un creodo), esta figura (**Fig. 5.8**) nos muestra también el linaje para cada tipo celular de GSEA diferenciando el compartimento hematopoyético del linaje de mesodermo. Podemos observar en la **Fig. 5.7** donde los tipos celulares asignados mayoritariamente pertenecen al compartimento hematopoyético.

Las asignaciones de los tipos celulares a los distintos creodos se muestran en la figura **Fig. 5.9** en la cual se observa mayor precisión a la hora asignar el linaje celular que a la hora de asignar el compartimento celular. En el caso de la identificación del linaje celular en 14 creodos no se consiguió ningún acierto, el resto de creodos acumularon principalmente entre 1 y 4 aciertos.

En la misma línea al realizar una 'asignación automática' a partir del número de aciertos (asignar el linaje o compartimento cuando hay 3 o más tipos celulares de un linaje o compartimento) observamos en torno al 48 % de aciertos a la hora de estimar el linaje celular y un total de 10 % de aciertos al asignar el linaje celular de forma automática en ambos casos (**Fig. 5.10**).

5.4. Discusión

Si nos atenemos a los resultados seleccionados por el algoritmo genético no se observan diferencias significativas entre diferentes combinaciones de genes-virtuales, esto puede significar que para el número de tipos celulares y creodos que manejamos en esta prueba, todas configuraciones responden parecido, si quisiéramos ver diferencias deberíamos probar números mayores, que implicarían que configuraciones mayores de genes-virtuales empezaría a ser un porcentaje grande de genes, respecto al total. Pero, el objetivo es buscar combinaciones mínimas de genes que expliquen el paisaje de Waddington de acuerdo a nuestra función objetivo y que sean lo suficientemente grandes para identificar la red GRN, la cual según la bibliografía (Fei et al., 2022) debe tratarse de un número considerable de genes, aunque es posible que sean ciertos genes esenciales para la toma de decisiones de las células durante la diferenciación celular y que por tanto sean determinantes en al elección de un linaje celular u otro (Fei et al., 2022).

Las combinaciones mayores de genes-virtuales implica menores tamaños del número de cromosomas-virtuales, lo que es el parámetro que más carga computacional genera, con la función objetivo actual. Posibles versiones futuras en este sentido podrían ir de la mano de implementar una paralelización 'multidimensional', ya que la actual implementación paraleliza las diferentes combinaciones de genes-virtuales y cromosomas-virtuales, una futura implementación con paralelización 'multidimensional' podría paralelizar a su vez diferentes combinaciones de genes-virtuales y

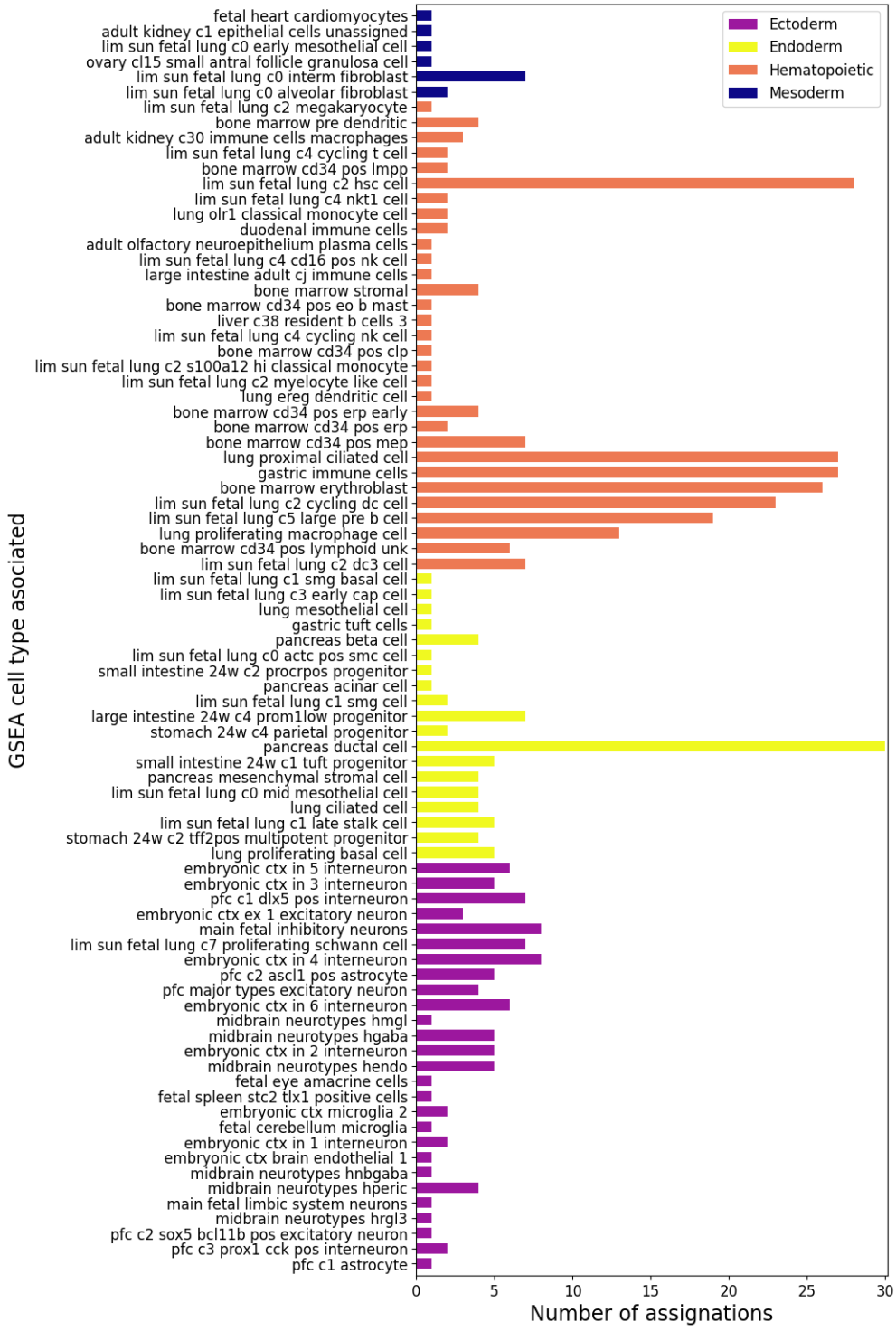


Figura 5.8: **Histograma de la distribución de los tipos celulares de GSEA emparejados con los creodos y la diferenciación de los atractores**. El color marca el linaje celular de cada tipo celular de GSEA, diferenciando *Hematopoietic* para los tipos celulares del linaje hematopoyético. En el eje x se muestran los nombres de los tipos celulares de GSEA y en el eje y se contabiliza las veces que ha sido asignado a los diferentes creodos.

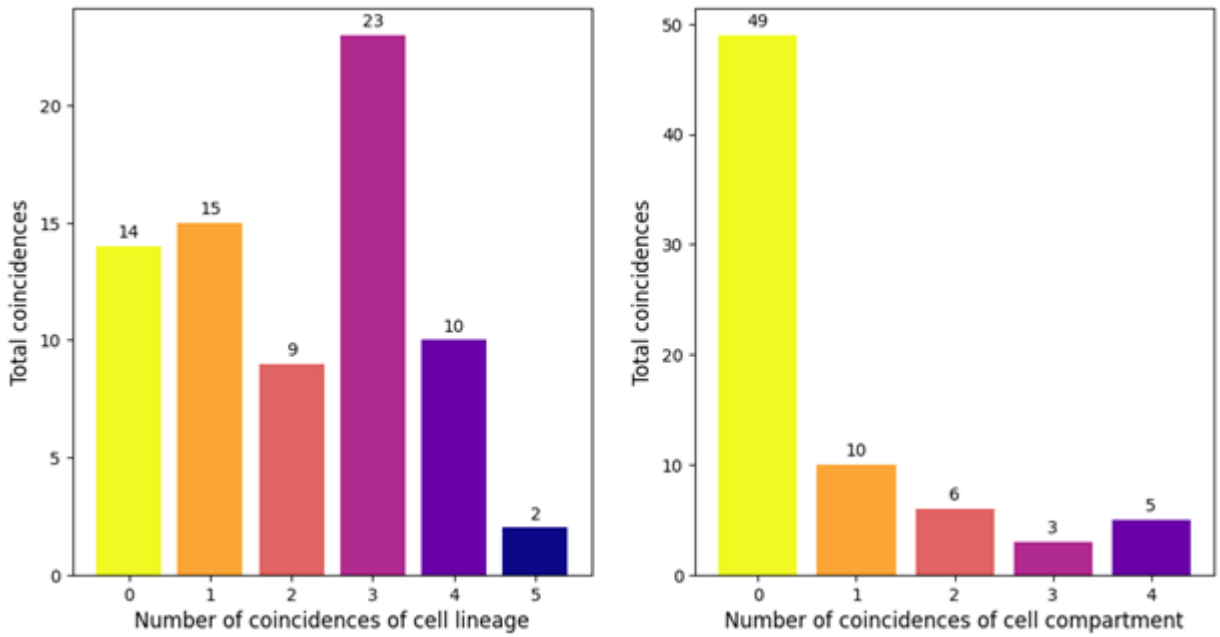


Figura 5.9: **Aciertos en el emparejamiento con tipos celulares de GSEA, para linaje celular (izquierda) y para compartimento celular (derecha)**. Los aciertos se contabilizan de 0 a 5 aciertos (eje x) para cada creodo y se muestra también el número de creodos agrupados por cantidad de aciertos (eje y).

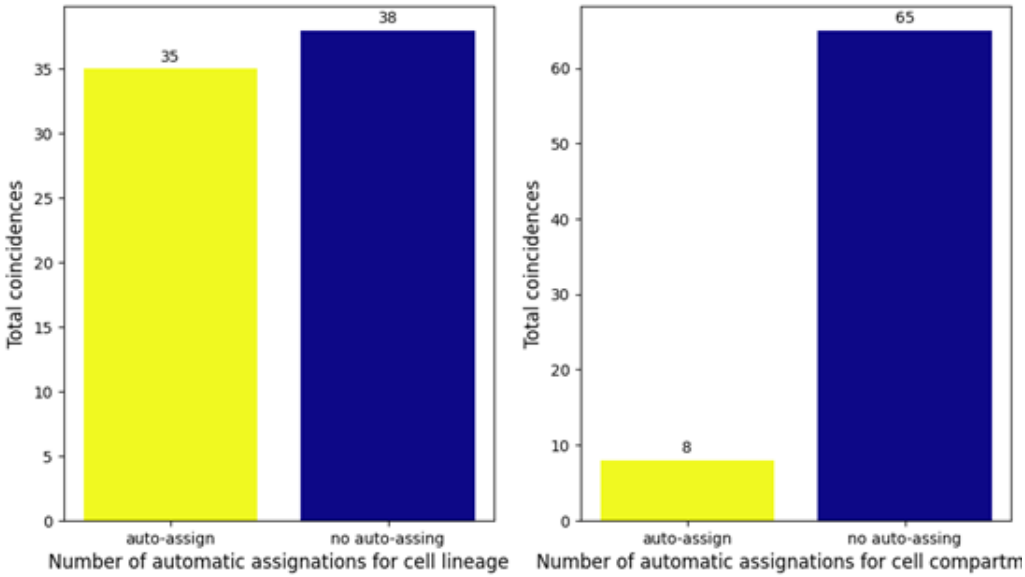


Figura 5.10: **Diagrama de barras de asignación automática de linaje y compartimento en función de los aciertos del emparejamiento con GSEA, en la izquierda se muestran los resultados para linaje celular y a la derecha los resultados para compartimento celular**. Aciertos y fallos de una asignación automática tras el emparejamiento con tipos celulares de GSEA para linaje celular y para compartimento celular. Las columnas de color amarillo representan aquellos tipos celulares que habría resultado asignados correctamente y en azul se representan los fallos

el cálculo de la puntuación de aptitud de conjuntos de cromosomas-virtuales, lo cual disminuiría la carga computacional de forma considerable y aumentaría la eficiencia del proceso. Aunque es posible que la construcción de este modelo de implementación requiriera la elaboración de nuevo software de paralelización.

Se observa una selección de términos de ontología de genes relacionados con el sistema inmunitario y/o con procesos hematopoyéticos, estos resultados son respaldados por el emparejamiento con tipos celulares de GSEA, donde se observa una selección de tipos celulares relacionados principalmente con el linaje hematopoyético/inmunitario. Esto nos lleva a un enriquecimiento del porcentaje de asignaciones del linaje del mesodermo, lo cual contrasta con nuestra distribución en linajes (**Fig. 3.5**). Es posible que también se deba a que nuestro linaje mayoritario se encuentra en torno al cigoto, mientras que en la base de datos de huella transcriptómica de tipos celulares de GSEA no se encuentras datos sobre tipos celulares encuadrados dentro del linaje cigoto.

A la hora de evaluar las asignaciones basándonos en los tipos celulares de GSEA, observamos que no existe una correlación entre los tipos celulares atractores de los creodos con términos de ontología de genes significativos y las asignaciones correctas de GSEA.

$$probabilidad\ 5\ errores = \left(\frac{x-1}{x}\right)^5 \qquad \text{(Ec. 5.4.1)}$$

$$probabilidad\ 5\ aciertos = \left(\frac{1}{x}\right)^5 \qquad \text{(Ec. 5.4.2)}$$

Donde x es el conjunto de posibles resultados que disponemos, 4 en caso de linaje celular (mesodermo, ectodermo, endodermo y cigoto) y 45 en caso de compartimento celular (**Fig. 5.1**).

Según el cálculo probabilístico (**Ec. 5.4** y **Ec. 5.4**), la asignación azar en el caso del linaje nos daría una probabilidad de $0{,}23$ de obtener 5 fallos y un $9{,}76E^{-4}$ de probabilidad de obtener 5 aciertos. Mientras, respecto al compartimento celular, nos encontramos con una probabilidad del $0{,}89$ de obtener 5 fallos y una probabilidad de $5{,}42E^{-9}$. En esencia, esto supondría 17 creodos con 0 coincidencias para el caso de la identificación de linaje celular y un total de 64 creodos con 0 aciertos para el mismo caso. Este cálculo probabilístico asumiría que la base de datos de GSEA dispone de huellas de tipos celulares de todos los linajes y de todos los compartimentos, sin grandes diferencias.

En ambos casos se obtienen valores con 0 aciertos por debajo de la asignación al azar, aunque estos siguen siendo una cantidad importante del total de casos. Es posible que estos resultados se den debido a que los datos públicos en la base de datos de huellas transcriptómicas de GSEA no dispongan de tipos celulares del linaje cigoto y es posible que tampoco contengan tipos celulares de todos los compartimentos.

Una identificación automática supondría identificar la mayoría de los tipos celulares como miembros de un determinado linaje o compartimento celular. En nuestro caso si seleccionamos las 5 mejores asignaciones para una identificación automática necesitaríamos que al menos 3 de las 5 asignaciones fuesen en un mismo linaje o compartimento celular (**Fig. 5.10**). En el caso del linaje celular la identificación automática acertaría aproximadamente la mitad de los creodos mientras que para el compartimento celular solo acertaría el 10 % de los creodos. Y aunque aún hay que mejorar el proceso, los resultados indican que con el aumento de las contribuciones a la base de datos de GSEA o con el uso de otras alternativas fruto de la iniciativa del Human Cell Atlas **(Rozenblatt-Rosen et al., 2017)** un paisaje de Waddington generado podría identificar correctamente el linaje y posiblemente el compartimento celular generando una herramienta útil en este sentido para la identificación y el descubrimiento de nuevos tipos celulares.

Capítulo 6

Discusión general

A nivel general, los objetivos del presente trabajo fueron orientados a la construcción de un paisaje de Waddington integral basado en datos transcriptómicos. Un paisaje de Waddington nos permite estudiar los mecanismos que gobiernan la diferenciación celular. Comprender estos mecanismos tendría una aplicación directa en el diseño de estrategias de reprogramación celular y para observar las alteraciones en el paisaje que provocan las muestras patológicas como medio para una mayor comprensión de los mecanismos patológicos. Con un paisaje de Waddington integral que tenga una alta representación podría servirnos para la identificación de muestras y el etiquetado automático de las mismas. La construcción de un paisaje de Waddington, no solo nos permite la identificación de muestras y el establecimiento de huellas genéticas, si no que podría ser una herramienta útil para la búsqueda de *House Keeping genes*, los cuales se definen como genes estables en las células bajo diferentes condiciones, incluso con la presencia de patologías (Joshi et al., 2022); (Tilli et al., 2016); (Thellin et al., 1999), cuya utilidad está ligada para la creación de controles internos para la interpretación de análisis de mRNA (Joshi et al., 2022); (Tilli et al., 2016).

La construcción del paisaje de Waddington se descompuso en tareas más pequeñas: **creación de una ontología de tipos celulares**, **selección, filtrado y normalización de los datos de GEO**, **selección de rasgos con algoritmo genético** y finalmente **validación y evaluación los resultados del algoritmo genético**.

El punto de partida de la construcción del paisaje de Waddington es la obtención de datos transcriptómicos presentes en repositorios públicos, en nuestro caso, en el repositorio de *GEO database*. Para obtener los correspondientes datos transcriptómicos, y establecer jerarquías de desarrollo entre ellos, se requiere de la existencia de estándares para las etiquetas de los diferentes

tipos celulares así como un estándar de las relaciones de desarrollo. La forma por la que se ha optado para ordenar y jerarquizar esta información es mediante la **creación de una ontología de tipos celulares** que recoja la información sobre la diferenciación celular. Esta ontología es fruto de la fusión de CELDA y LifeMap, la cual es ampliable gracias al software de *alineamiento* y *merging* FOntCell que desarrollamos (Cabau-Laporta et al., 2021). Esta ontología supone un consenso entre las ontologías de LifeMap y CELDA y es el estándar que hemos utilizado en los siguientes pasos.

A continuación, se ha creado una **herramienta para la selección, filtrado y normalización de los datos alojados en la *web* de GEO.** Futuras versiones y actualizaciones de la web podrían dejar obsoleta la actual versión de la herramienta de *scrap web* y deberían introducirse cambios a esta herramienta que permitiera readaptarla a los posibles cambios en la *web* de GEO, u otras *webs*, y mantener la lógica de los algoritmos de búsqueda y etiquetado masivo de datos. En este punto cabe destacar que, aunque funcional, el alineamiento de cadenas de caracteres mediante la teoría de 'similitud' de Levenshtein (Levenshtein, 1966) nos permite comparar la similitud y buscar coincidencias entre dos cadenas de caracteres.

En este punto, se ha obtenido una serie de datos transcriptómicos de diferentes tipos celulares enmarcados dentro de las relaciones de desarrollo que nos proporciona la ontología de CELDA+LifeMap. Estos datos transcriptómicos parten de diferentes trabajos y como se ha mencionado anteriormente, el principal problema de utilizar datos de diferentes trabajos es el conocido 'efecto de *batch*' o *batch effect*; (Johnson, Li, 2007), los datos de un mismo trabajo sean más similares entre sí que con datos del mismo tipo perteneciente a otro trabajo. Con el objetivo de reducir el *batch effect* se ha utilizado la herramienta ComBat (Johnson, Li, 2007). Es posible que, en el futuro, y con el uso de otro tipo de tecnologías, surjan nuevas herramientas que ayuden a reducir el *batch effect* de forma más eficaz.

Con el paso del tiempo nuevos datos serán publicados. Esto facilitará la labor de etiquetado y obtención de datos, permitiéndonos ser más selectivos a la hora de eliminar muestras con las que tengamos dudas sin la contrapartida de perder representación. Una mayor representación de diferentes tipos celulares provocaría un mejor funcionamiento a la hora de utilizar el software de eliminación del *batch effect* utilizando las herramientas que disponemos en la actualidad. También, nos permitirá obtener mayor representación de todos los procesos, desde los estadios más indiferenciados a los diferentes atractores del paisaje de Waddington.

Con los datos transcriptómicos etiquetados y normalizados, es necesario realizar una selección de rasgos para la construcción del paisaje de Waddington y se ha abordado como un problema de optimización. Por ello, que se decidió el uso de un **algoritmo genético** para la selección de conjuntos de genes que estuvieran involucrados en la toma de decisiones de la diferenciación

celular en el contexto de un paisaje de Waddington.

La elaboración del software del algoritmo genético nos permite utilizarlo para otras aplicaciones que puedan plantearse como un problema de optimización. La parte más relevante para el algoritmo genético es la función objetivo. La función objetivo es en esencia la expresión matemática de la hipótesis de construcción del paisaje de Waddington que hemos implementado en el presente trabajo. En nuestra interpretación del paisaje de Waddington, entendemos que las relaciones entre elementos indiferenciados y atractores podría darse en cuanto a conjuntos de genes que tuvieran relaciones decrecientes respecto a su expresión.

Los métodos para la **validación y evaluación los resultados del algoritmo genético** nos permiten extraer información del conjunto de genes resultantes y ver con qué términos de función biológica correlacionan en un test GO.

Creación de ontología de tipos celulares

El etiquetado de los diferentes tipos celulares presenta pequeñas discrepancias en cuanto al consenso de la nomenclatura, y en algunos casos de la jerarquía de diferenciación, entre diferentes ontologías (Edgar et al., 2013) (Seltmann et al., 2013). Por ello se desarrolló FOntCell (Cabau-Laporta et al., 2021). La idea principal en el desarrollo de FOntCell es la de fusionar dos ontologías (CELDA y LifeMap) intentando conservar el máximo posible de tipos celulares y sus diferentes sinónimos. Tal y como se ha comentado, al no existir un consenso claro en las ontologías de origen genera que la ontología resultante tenga tipos celulares que comparten sinónimos con otros tipos celulares generando una especie de duplicidad de tipos celulares si atendemos a los sinónimos. Estos sesgos son difíciles de eliminar por FOntCell y requeriría una actualización que incorporara un algoritmo específico para eliminarlo. Por otra parte, si un tipo celular es etiquetado con unos sinónimos por las ontologías y existe discrepancia es posible que esos tipos celulares no estén definidos completamente en la bibliografía por lo que un algoritmo que solucione estas discrepancias es posible que esté introduciendo un nuevo sesgo.

Centrándonos en el aspecto del alineamiento de ontologías, FOntCell consiguió unos resultados competitivos para la OAEI (Capítulo 2) pero la iniciativa del alineamiento de ontologías es un campo en constante expansión y es posible que eventualmente aparezcan algoritmos de alineamiento mejorados que superen a los actuales que incorpora FOntCell.

Herramientas para la selección, filtrado y normalización de los datos alojados en la *web* de GEO.

El conjunto de algoritmos implicados en la selección, filtrado y normalización de datos es una

parte clave en la construcción del paisaje de Waddington, ya que preparan las muestras alojadas dándoles una etiqueta para el después procesadas por el algoritmo genético. El principal punto positivo es que muchos de estos algoritmos pueden ser adaptados para hacer la misma función en otro tipo de datos ya que exclusivamente analizan los metadatos alojados en la web.

En la actual versión, para la plataforma GPL570 consiguen procesar el 8 % de las muestras alojadas en esta plataforma. Una versión futura debería obtener un mayor número de tipos celulares diferentes, sobretodo de tipos celulares intermediarios dentro de los creodos. Esta ampliación del número de muestras viene de la mano de un mejor análisis de los metadatos de las muestras alojadas. Los metadatos de los diferentes GSEs y GSMs son evaluados por nuestros algoritmos mediante la ecuación de Levenshtein (Levenshtein, 1966). En la actualidad, la aplicación de actuales modelos de lenguaje (Brown et al., 2020) podría ser útil en la identificación de tipos celulares a partir de sus metadatos. El uso de los modelos de lenguaje, nos permitiría generar una base de datos más fiable que descarte en la medida de lo posible tipos celulares que sean patológicos y seleccione exclusivamente los controles. La construcción del paisaje de Waddington nos ayudaría mediante un proceso iterativo en el establecimiento de huellas genéticas concretas de cada tipo celular y facilitar así un control adicional en la base de datos, para confirmar que tanto el nombre asignado como la huella transcriptómica coinciden, lo cual ahorraría mucho tiempo de evaluación al realizarlo con un software automático, además de la posibilidad de identificar aquellas muestras cuyo etiquetado no es posible identificarlo fácilmente (uso de siglas, acotaciones, etc).

Aunque nuestras herramientas de selección y filtrado de muestras discriminan ciertas etiquetas que disponen una serie de términos prohibidos (tal y como se muestran en la **??**) no pueden garantizar que se añadan muestras de tipos celulares afectados por distintas patologías. Ya que en ocasiones los campos de texto más descriptivos e individuales de un tipo celular no reúnen información necesaria y le es imposible a nuestro algoritmo distinguir las muestras con afecciones de las muestras control. En esta línea, lo ideal sería la producción de datos específicos para este análisis o el reciclado de trabajos sobre diferenciación celular. La tecnología single-cell sería también interesante para un proyecto de este estilo ya nos permitiría tener muestras más concretas de tipos celulares concretos (Zappia et al., 2018).

La normalización de los datos vendrá en función del tipo de dato utilizado para la construcción del paisaje de Waddington. Además de los protocolos existentes dependiendo de la plataforma utilizada, si se utilizan datos de diferentes trabajos es necesario utilizar software para la corrección del *batch effect*, lo cual es una línea de investigación en sí mismo y es esperable la aparición de nuevos algoritmos para la corrección del *batch effect* en los próximos años sobretodo aplicado a las nuevas tecnologías ómicas. Una alternativa sería la de utilizar la llamada lógica difusa. La lógica difusa supondría asumir categorías de expresión (Cintula et al., 2021), por ejemplo, en

una lógica difusa con 3 categorías podríamos asumir: alto, medio y bajo. La principal ventaja de estos sistemas supondría el poder analizar muestras transcriptómicas analizadas con diferentes tecnologías (Sehgal et al., 2006), las cuales a día de hoy no pueden ser comparables unas con otras, siempre y cuando el número de categorías no sea demasiado elevado. Aunque el uso de la lógica difusa requiere el desarrollo de una serie de algoritmos para 'transcribir' cada una de las tecnologías que se quieran implementar y es posible que se alteren los resultados lo suficiente de forma que se pierda parte de la información.

Algoritmo genético

El algoritmo genético que se ha desarrollado permite encontrar los genes cuya expresión genética satisface en mayor medida la función objetivo a partir de una información ontológica y unos datos transcriptómicos asociados a los mismos obtenidos con anterioridad. Aun así, respecto al apartado técnico del funcionamiento del algoritmo genético pueden incorporarse algunas mejoras. Las mejoras que más pueden afectar a la implementación son aquellas relacionadas con la paralelización del software. Paralelizar el cálculo de los diferentes cromosomas-virtuales reduciría el tiempo de ejecución del algoritmo. Otra mejora técnica a incorporar sería un sistema de mutación, típico de los algoritmos genéticos (Holland, 1992), donde al pasar de una generación a la siguiente cada uno de los genes-virtuales tiene una posibilidad de mutación (m), proceso por el cual el gen-virtual mutado se convertiría en un gen-virtual del *pool* inicial de genes-virtuales incluyendo aquellos genes-virtuales descartados. La mutación nos permitiría una tasa de recursividad adicional, asegurando una mejor convergencia, aunque presumiblemente podría aumentar la carga computacional.

Previo al algoritmo genético se seleccionan los genes con mayor variabilidad en el conjunto de muestras basándonos en la puntuación de cada gen en un análisis de la varianza. En este punto podría ser interesante realizar un análisis basado en el GeneGini (Joshi et al., 2022) el cual se basa en el coeficiente Gini para medir la desigualdad entre grupos, esto nos permite calcular la inequidad en la expresión entre las muestras, dándonos un valor entre 0 y 1 a cada uno de los genes, siendo los más estables entre las células cercanos a 0 y los más variables cercanos a 1 (Joshi et al., 2022), esto puede ayudarnos como métrica para realizar un filtrado previo al algoritmo genético, desechando aquellos genes con menor coeficiente GeneGini de la misma forma que desechamos los menos variables. Es posible que esta métrica nos dé un conjunto más reducido y nos seleccione en mayor medida el grupo de genes que más información guarda entre las distintas muestras que disponemos.

La función objetivo es la expresión matemática que queremos optimizar. Esta expresión matemática es en cierta forma, la expresión matemática del paisaje de Waddington y está relacionada con la hipótesis del paisaje de Waddington. En nuestro caso hemos utilizado 3 funciones objetivo,

DB, DC y DEC. DB y DC son funciones que podemos etiquetar como 'globales' ya que buscan un conjunto de genes que expliquen la diferenciación en todo el organismo. Esto nos genera principalmente dos escenarios desfavorables:

Por un lado, evaluar todos los tipos celulares en su conjunto supone de entrada asumir que la diferenciación a lo largo de todos los tipos celulares es una aproximación reduccionista del paisaje de Waddington al asumir una única función de altura para todo el conjunto de relaciones de diferenciación que integren nuestro paisaje de Waddington (Sáez et al., 2022). En nuestro caso, al asumir mecanismos intrínsecos de cada creodo de forma independiente asumimos de forma tácita la toma de decisiones celular, la cual podría extraerse de los resultados mediante un estudio comparado de dos creodos, aunque es posible que para desentrañar los mecanismos de toma de decisiones y comprender e identificar con precisión los mecanismos de dediferenciación sea necesario tener en cuenta más tipos de datos ómicos (Wang et al., 2022); (Li et al., 2022); (Lord, Nixon, 2020); (Tyurin-Kuzmin et al., 2020); (Gibb et al., 2020) en la construcción del paisaje de Waddington, e incluso tender a paisajes de Waddington dinámicos (Sáez et al., 2022)

Por otro lado, una métrica basada únicamente en el DB puede verse fácilmente alterada por el ruido que el propio análisis genere cuando nos encontramos genes con transcripción similar. En nuestro caso, optamos por el uso del DEC, ya que fuerza la selección de relaciones de decrecimiento de mayor envergadura y evita la selección de diferencias causada por el ruido.

Es por ello que para buscar un ajuste de la función de validación a los datos para la construcción del paisaje de Waddington se implementó la separación por creodos y, adicionalmente, se dio más peso a las relaciones de decrecimiento seleccionando aquellas que describieran con mayor ajuste a una curva de decaimiento exponencial a lo largo de la expresión del conjunto de genes que se estudien para cada creodo. El DEC, como ya se ha comentado con anterioridad, nos permite seleccionar características en decrecimiento 'brusco' de modo que evitemos la selección del ruido de la propia tecnología de análisis.

Combinando la evaluación de cada creodo por separado y la relación de DEC obtenemos, en la prueba de concepto, resultados estadísticamente significativos en el conjunto de genes seleccionados para cada uno de los creodos en un análisis de enriquecimiento de términos de ontología de genes, lo que supone una mejoría en los resultados respecto a las otras alternativas. Sin embargo, a la hora de evaluar el conjunto de datos del paisaje de Waddington se consiguen datos estadísticamente significativos para el análisis de términos de la ontología de genes para aproximadamente la mitad de los creodos.

Nuevas hipótesis de construcción del paisaje de Waddington requerirán otras formulaciones de la función objetivo respecto a los datos transcriptómicos o quizás, otro tipo de datos ómicos, como

por ejemplo el metaboloma. Por ejemplo, podría construirse un paisaje de Waddington inverso al nuestro, en el que se seleccionen aquellos conjuntos de genes cuya transcripción sea creciente desde el inicio de los creodos hasta los distintos atractores, por ejemplo. Aunque esta idea es interesante, cabe la posibilidad de que aún no sea posible implementarla debido a la dispersión actual de las muestras (algunos de los creodos no tienen demasiados intermediarios entre los estadios cigóticos y el atractor final).

Métodos de validación y evaluación de los resultados del algoritmo genético

Los resultados del algoritmo genético con la función objetivo DEC son un conjunto de genes que satisfacen un decaimiento exponencial en su expresión desde los nodos iniciales a los diferentes atractores, los métodos de validación y evaluación de estos resultados pretende observar el tipo de genes, en conjunto, que han sido seleccionados por el algoritmo genético. La hipótesis que subyace a nuestras funciones de aptitud es que existe un conjunto de genes cuya transcripción decrece a lo largo de la diferenciación, es decir, nuestro modelo selecciona aquellos genes que en estadios iniciales tienen una mayor expresión que en los momentos finales en los cuales se van 'apagando', es decir, se pretende buscar los genes 'disparadores'.

En el caso del análisis GO, se buscan patrones significativos de función de los genes seleccionados, para así encontrar posibles funciones relacionadas con la diferenciación. Versiones futuras del paisaje de Waddington podrían localizarse en conjuntos más pequeños y observar si los genes seleccionados correlacionan con las funciones de diferenciación encontradas en la bibliografía.

El análisis de GSEA, es un análisis construido para el presente trabajo. En este análisis comparamos los genes resultantes con la huella genética de cada uno de los tipos celulares registrados en la web de GSEA, la principal hipótesis que subyace a este análisis que los genes seleccionados para un determinado creodo tengan una coincidencia mayor para tipos celulares de ese linaje y concretando más, para ese compartimento. Es posible que un paisaje de Waddington futuro, en el que se use tecnología single-cell genere mejores resultados que en la actual versión ya que la base de datos de GSEA está construida con datos single-cell. También, al estar comparando los resultados producidos por una función objetivo basada en un DEC con las huellas transcriptómicas de tipos celulares en muchos casos diferenciados es posible que obtuviéramos mejores coincidencias si realizáramos un paisaje de Waddington inverso, es decir, utilizar funciones de aptitud que maximicen genes con tendencia creciente a lo largo de los creodos, lo cual queda pendiente para futuras versiones y podría completar con nueva información el actual paisaje de Waddington.

Resultados biológicos

Nuestro algoritmo genético ha sido diseñado para seleccionar aquellos genes cuya transcripción es más influyente en la toma de decisión celular, principalmente en los estadios iniciales. Para extraer información del conjunto de genes seleccionados por el algoritmo genético se aplicaron dos test, uno basado en GO y otro en comparaciones con las huellas genéticas de la base de datos de GSEA.

Observando la comparación con las huellas genéticas de GSEA como identificador de tipos celulares, nos encontramos con que el caso de identificación del linaje es más favorable. Entendemos que la identificación del compartimento celular es 'hilar demasiado fino' para el volumen de datos que disponemos, ya que nuestra muestra original no dispone de muchos tipos celulares intermediarios o próximos a los compartimentos de los atractores, además de que la propia función de validación satisface genes con gran expresión en los estadios iniciales por lo que se espera que sea más efectivo en la detección de linajes (ya que sucede en estadios más tempranos) que en detección de compartimentos, lo cual encaja con los resultados obtenidos. Otro posible sesgo puede venir de la mano de las huellas de tipos celulares alojados en GSEA, ya que hay datos de todos los compartimentos (a excepción de cigoto) pero no tiene por qué haber representación de todos los compartimentos en la actualidad. Esto refuerza la idea de que los resultados de linaje en última instancia son más fiables.

Aun así, los resultados obtenidos del algoritmo genético son suficientes como para sacar lecturas e interpretaciones de los mismos en el estado actual, donde hemos podido observar tanto en el GO como en el análisis de emparejamiento de tipos celulares de GSEA, un enriquecimiento tanto en términos relacionados con procesos inmunológicos/hematopoyéticos como un emparejamiento principalmente con tipos celulares del sistema inmune/de linaje hematopoyético. Adicionalmente, en el GO se ha observado también un enriquecimiento en términos relacionados con el metabolismo celular, donde incluimos procesos de catabolismo y anabolismo.

Contextualizando con los resultados de otros trabajos encontramos coincidencias con los resultados expuestos por Fei et al. (2022), donde en su construcción del paisaje de Waddington el gen *XBP1* juega un papel central para la toma de decisiones del destino celular (Fei et al., 2022). En nuestra matriz transcriptómica el gen *XBP1* no fue seleccionado entre los genes con mayor variabilidad entre las muestras, por lo que no fue incluido en los análisis posteriores basados en el algoritmo genético, pero cabe destacar los términos de la ontología de genes asociados al gen *XBP1* se relacionan principalmente con procesos de diferenciación del sistema inmune (Muralidharan, Mandrekar, 2013); (Ono et al., 1991), rasgos también seleccionado en nuestro propio análisis.

XBP1 comparte también un rasgo junto con otros genes que comparten términos de la ontología de genes con otros del sistema inmune, y es que muchos de los genes asociados al sistema

inmune se encuentran expresados en células no pertenecientes al mismo bajo circunstancias de estrés (donde juega un papel importante en la respuesta al estrés de Retículo Endoplasmático relacionado con el cáncer) (Muralidharan, Mandrekar, 2013); (Fulda et al., 2010); (Ono et al., 1991). Y han sido observados en altas expresiones durante la diferenciación Han et al. (2020).

El enriquecimiento en procesos metabólicos del GO se observa en procesos relacionados con la diferenciación de distintos tipos celulares en la bibliografía. En las células-T el metabolismo puede producir cambios epigenéticos y por tanto influyen la diferenciación de células-T naive a células-T de memoria o efectoras (Li et al., 2022). Estos cambios pueden venir de parte de la proximidad a las células sinápticas inmunes que proveen de metabolitos o bien debido a cambios en el ambiente producidos por algún tipo de estrés (infección o inflamación) (Li et al., 2022). Existen numerosos ejemplos donde el metabolismo influye en la diferenciación celular como es el proceso de la espermatogénesis (Lord, Nixon, 2020), los myofibroblastos (Gibb et al., 2020) y en el desarrollo y diferenciación de los tipos celulares de riñón (Wang et al., 2022). En líneas generales, encontramos que los metabolitos producen cambios epigenéticos que llevan en ocasiones a la diferenciación celular. Es importante señalar que el metabolismo cambia a lo largo de la diferenciación así que parece ser que en cierta forma el metabolismo puede producir cambios en la diferenciación y también la diferenciación produce cambios en el metabolismo.

Esto nos genera múltiples lecturas y posibles interpretaciones de este fenómeno:

- La opción más sencilla, atendiendo a los resultados de la comparativa de huellas transcriptímicas de GSEA, parece ser que exista un enriquecimiento de tipos celulares del sistema inmune en el repositorio de GSEA, donde existe un total de 830 muestras de tipos celulares. Esto podría generar la situación en la que si el sistema no consigue identificar correctamente un tipo celular asignaría cualquier otro y si coincide con que los tipos celulares del linaje hematopoyético son los más comunes lo más posible sería que se anotara como uno de ellos. Aunque esta opción no explicaría el enriquecimiento de términos relacionados con la inmunidad del GO.

- Por un lado, teniendo en cuenta que nuestros datos pueden sufrir de un 'efecto centro' a la hora de seleccionar genes con alto nivel transcriptómico en los estadios iniciales y una disminución exponencial hacia los atractores, es posible que el estrés producido a estas células durante su análisis (sobre todo si son de naturaleza patológica) genere la sobreexpresión de genes que se asocian al sistema inmune, o al menos que tienen términos de la ontología de genes asociados a procesos del sistema inmune. Existen evidencias de que células sometidas a diferente estrés expresan genes del sistema inmune (Mukherjee et al., 2023); (Muralidharan, Mandrekar, 2013) y es posible que el estrés produzca también cambios metabólicos (Gibb et al., 2020).

- Por otro lado, es posible que los genes asociados con el sistema inmune y del metabolismo jueguen un papel determinante en la diferenciación celular y en la red genética de regulación (GRN) en los procesos de selección de destinos celulares, visión que coincide con los datos recogidos de estudios previos sobre diferenciación (Fei et al., 2022); (Han et al., 2020) y por estudios de diferenciación y metabolismo (Wang et al., 2022). Se ha encontrado expresión de genes del sistema inmune por parte de células no inmunes durante la diferenciación, tanto en células adultas como de células fetales a células adultas (Han et al., 2020) e incluso se ha observado que XBP1 juega un papel clave en la decisión del linaje celular (Fei et al., 2022). Y también, se ha observado que el metabolismo puede ser un conductor de la diferenciación genética y por tanto podría utilizarse como control epigenético (Wang et al., 2022).

- Y por otro lado, podría ser que el estrés celular genere una especie de dediferenciación celular con el objetivo de mantenerse a la 'espera' de que el estímulo de estrés desaparezca y ocupar nuevos nichos en el organismo que hayan quedado 'vacíos' como consecuencia de ese estrés, es decir, que en presencia de estrés celular una célula activa una serie de mecanismos de dediferenciación y diferenciación celular ligados también al sistema inmune y procesos metabólicos con el objetivo de adaptarse, resistir y reparar los daños causados por ese estrés celular.

Otros aspectos que es discutido en la bibliografía actual respecto a la implementación de la metáfora del paisaje de Waddington discuten el tipo de dato con el que construir este paisaje de Waddington (Loison, 2022); (Bizzarri et al., 2020). Consideramos la transcriptómica uno de los puntos claves del paisaje de Waddington, tal y como él creía (Loison, 2022);(Noble, 2015) pero es posible que esto sea una simplificación de los mecanismos celulares implícitos para la toma de decisiones de la diferenciación celular (Sáez et al., 2022) y hace necesario que para un paisaje de Waddington más preciso sea necesario tener en cuenta otros datos ómicos de los diferentes estadios celulares (Bizzarri et al., 2020);(Sáez et al., 2022);(Loison, 2022). En el trabajo y metáfora original de Conrad H. Waddington ya se habla de la 'plasticidad' del paisaje de Waddington, el cual se altera debido a lo que él llamó 'epigenética' para dar lugar a nuevos creodos (Noble, 2015);(Loison, 2022), en la actualidad los paisajes de Waddington dinámicos se están entendiendo con los términos de 'bifurcaciones locales' y 'bifurcaciones globales', lo cual denomina a las regiones del paisaje de Waddington haciendo alusión al tamaño del 'terreno' que alteran del paisaje de Waddington (Sáez et al., 2022). En esta nueva familia de paisajes de Waddington, se habla de paisajes que tengan en cuenta principalmente relaciones epigenéticas (que podrían entenderse como una influencia del entorno) (Bizzarri et al., 2020) e interacciones célula-célula, generando un paisaje de Waddington dinámico (Sáez et al., 2022).

Un paisaje de Waddington dinámico podría ayudar al entendimiento de la toma de decisiones de

destino de las células y por tanto ayudar al establecimiento de estrategias concretas para alterar esta toma de decisiones. Presentamos una serie de herramientas para la construcción de un paisaje de Waddington, planteándonos la selección de los elementos más relevantes como un problema de optimización resuelto con algoritmos genéticos.

En definitiva, se ha construido un paisaje de Waddington estático, con las herramientas aquí desarrolladas, que nos permite estudiar los mecanismos intrínsecos al desarrollo celular en el ámbito de la transcriptómica. Futuras versiones del paisaje de Waddington, que incluyan aspectos dinámicos del mismo así como otros datos ómicos, muestras de *single-cell* y un mayor *coverage*, podrían ser una herramienta útil para anotar muestras *single-cell* de forma automática, descubrir nuevos estadios intermediarios (Chai, 2022) y comprender los mecanismos de la toma de decisiones de destino celular con precisión (Sáez et al., 2022) y por tanto, poder elaborar estrategias de reprogramación celular.

Capítulo 7

Conclusiones

A la luz de los resultados obtenidos y su posterior discusión podemos concluir:

- Se ha creado una herramienta de alineamiento y *merging* de ontologías, la cual nos construir y computar las relaciones de diferenciación de los tipos celulares.
- Mediante el uso de las herramientas de *scrap web* desarrolladas para este trabajo, se ha obtenido una base de datos de tipo celular utilizando repositorios públicos, lo cual nos ha permitido etiquetar el 8 % de los tipos celulares de GEO para la plataforma GPL570 de *Affymetrix*.
- Se ha creado un algoritmo genético con diferentes aproximaciones de la función objetivo: DB, DC y DEC, con la finalidad de seleccionar un conjunto de genes para la construcción del paisaje de Waddington.
- Se ha construido un paisaje de Waddington utilizando los datos transcriptómicos obtenidos mediante las herramientas *scrap web* en la plataforma GPL570 de *Affymetrix* y seleccionando la 'energía quasi-potencial' mediante la adición de la expresión transcriptómica de conjuntos de genes seleccionados mediante el algoritmo genético con la función objetivo DEC.
- Se ha observado que el enriquecimiento de términos de ontología de genes en nuestro modelo del paisaje de Waddington, utilizando el ajuste de DEC en el algoritmo genético, nos generan términos significativos relacionados con funciones inmunológicas y metabólicas.

Bibliografía

Altman D. G., Bland J. M. Statistics notes: The normal distribution // BMJ. 1995. 310, 6975. 298–298.

Antoniou G., Harmelen F. Web ontology language: OWL // Handbook on Ontologies. 2009. 91–110.

Ardini-Poleske M. E., Clark C. R. F.and Ansong, Carson J. P., Corley R. A., Deutsch G. H., Hagood J. S., Kaminski N., Mariani T. J., Potter S. S., Pryhuber G. S., Warburton D., Whitsett J. A., Palmer S. M., Ambalavanan N., Consortium LungMAP. LungMAP: The Molecular Atlas of Lung Development Program // American journal of physiology. Lung cellular and molecular physiology. 2017. 5, 313. L733–L740.

Bard J., Rhee S. Y., Ashburner . An ontology for cell types // Genome Biology. 2005. 6. R21.

Bizzarri M., Giuliani A., Minini M., Monti N., Cucina A. Constraints shape cell function and morphology by canalizing the developmental path along the waddington's landscape // BioEssays. 2020. 42, 4. 1900108.

Blondel V.D., Gajardo A.v, Heymans M., Senellart P., Van Dooren P. A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching // SIAM Review. 2004. 46, 4. 647–666.

Boldog E., Bakken Trygve E., Hodge R. D., Novotny M., Aevermann B. D., Baka J., Bordé S., Close J. L., Diez-Fuertes F., Ding S., Farago N., Kocsis A., Kovacs B., Maltzer Z., McCorrison J., Miller J., Molnar G., Olah G., Ozsvar A., Rozsa M., Shehata S., Smith K., Sunkin S., Tran D., Venepally P., Wall A., Puskas L., Barzo P., Steemers F., Schork N., Scheuermann R., Lasken R., Lein E., Tamas G. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type // Nature Neuroscience. 2018. 21, 9. 1185–1195.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 33. 2020. 1877–1901.

Busse J., Humm B., Lubbert C., Moelter F., Reibold A., Rewald M., Schluter .V, Seiler B., Tegtmeier E., Zeh T. Actually, What Does "Ontology" Mean? A Term Coined by Philosophy in the Light of Different Scientific Disciplines. Journal of Computing and Information Technology // CIT. 2015. 23, 1. 29–41.

Cabau-Laporta J., Ascensión A. M., Arrospide-Elgarresta M., Gerovska D., Araúzo-Bravo M. J. Font-cell: Fusion of ontologies of cells // Frontiers in Cell and Developmental Biology. 2021. 9.

Cappa F., Oriani R., Peruffo E., McCarthy I. Big data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety, and veracity on firm performance* // Journal of Product Innovation Management. 2020. 38, 1. 49–67.

Chai R. C. Single-cell RNA sequencing: Unravelling the bone one cell at a time // Current Osteoporosis Reports. 2022. 20, 5. 356–362.

Cintula P., Fermüller C. G., Noguera C. Fuzzy logic. Nov 2021.

Dzierzak E., Bigas A. Blood development: Hematopoietic stem cell dependence and independence // Cell Stem Cell. 2018. 22, 5. 639–651.

Edgar R., Domrachev M., Lash A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository // Nucleic Acids Research. 1. 2002. 30. 207–210.

Edgar R., Mazor Y., Rinon A., Blumenthal J., Golan Y., Buzhor E., Livnat I., Ben-Ari S., Lieder I., Shitrit A., Gilboa Y., Ben-Yehudah A., Edri O., Shraga N., Bogoch Y., Leshansky L., Aharoni S., West M.D., Warshawsky D., Shtrichman R. LifeMap Discovery™: the embryonic development, stem cells, and regenerative medicine research portal // PLoS One. 2013. 8, 7.

Fard A. T., Ragan M. A. Quantitative modelling of the Waddington Epigenetic Landscape // Computational Stem Cell Biology. 2019. 157–171.

Fard A. T., Srihari S., Mar J. C., Ragan M. A. Not just a colourful metaphor: Modelling the landscape of cellular development using Hopfield Networks // NPJ Systems Biology and Applications. 2016. 2, 1.

Fei L., Chen H., Ma L., E W., Wang X. R.and Fang, Zhou Z., Sun H., Wang J., Jiang M., Wang X., Yu C., Mei Y., Jia D., Zhang T., Han X., Guo G. Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development // Nature Genetics. 2022. 54, 7. 1051–1061.

Fulda S., Gorman A. M., Hori O., Samali A. Cellular stress responses: Cell survival and cell death // International Journal of Cell Biology. 2010. 2010. 1–23.

Gerovska D., Araúzo-Bravo M. J. Does mouse embryo primordial germ cell activation start before implantation as suggested by single-cell transcriptomics dynamics? // Mol. Hum. Reprod. 2016. 22. 208–225.

Gerovska D., Araúzo-Bravo M. J. Computational analysis of single-cell transcriptomics data elucidates the stabilization of Oct4 expression in the E3.25 mouse preimplantation embryo // Sci. Rep. 2019. 9. 8930.

Gibb A. A., Lazaropoulos M. P., Elrod J. W. Myofibroblasts and fibrosis // Circulation Research. 2020. 127, 3. 427–447.

Grindberg R. V., Yee-Greenbaum J. L., McConnell M. J., Novotny M., O'Shaughnessy A. L., Lambert G. M., Araúzo-Bravo M. J., Lee J., Fishman M., Robbins G. E., Lin X., Benepally P., Badger J. H., Galbraith D. W., Gage F. H., Lasken R. S. RNA-sequencing from single nuclei // Proceedings of the National Academy of Sciences. 2013. 110, 49. 19802–19807.

Grün D., Oudenaarden A. Design and analysis of single-cell sequencing experiments // Cell. 2015. 163, 4. 799–810.

Han X., Zhou Z., Fei L., Sun H., Wang R., Chen Y., Chen H., Wang J., Tang H., Ge W., Zhou Y., Ye F., Jiang M., Wu J., Xiao Y., Jia X., Zhang T., Ma X., Zhang Q., Bai X., Lai S., Yu C., Zhu L., Lin R., Gao Y., Wang M., Wu Y., Zhang J., Zhan R., Zhu S., Hu H., Wang C., Chen M., Huang H., Liang T., Chen J., Wang W., Zhang D., Guo G. Construction of a human cell landscape at single-cell level // Nature. 2020. 581, 7808. 303–309.

Heins N., Malatesta P., Cecconi F., Nakafuku M., Tucker K. L., Hack Mi. A., Chapouton P., Barde Y. A., Götz M. Glial cells generate neurons: The role of the transcription factor PAX6 // Nature Neuroscience. 2002. 5, 4. 308–315.

Holland J. H. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and Artificial Intelligence // The MIT Press. 1992.

Huang S. Reprogramming cell fates: Reconciling rarity with robustness // BioEssays. 2009. 31, 5. 546–560.

Huang S., Guo Y., May G., Enver T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells // Developmental Biology. 2007. 305, 2. 695–713.

Hwang B., Lee J.H., Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines // Experimental & Molecular Medicine. 2018. 50, 96.

Ingram N. Waddington, Holmyard and alchemy: Perspectives on the epigenetic landscape // Endeavour. 2019. 43, 3. 100690.

Irizarry R. A., Hobbs B., Collin F., Beazer-Barclay Y. D., Antonellis K.J., Scherf U., Speed T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data // Biostatistics (Oxford, England). 2003. 4, 2. 249–264.

Johnson W. E., Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods // Biostatistics (Oxford, England). 2007. 8, 1. 118–127.

Joshi C. J., Ke W., Drangowska-Way A., O'Rourke E. J., Lewis N. E. What are housekeeping genes? // PLOS Computational Biology. 2022. 18, 7.

Ladewig J., Koch P., Brüstle O. Leveling waddington: The emergence of direct programming and the loss of Cell Fate Hierarchies // Nature Reviews Molecular Cell Biology. 2013. 14, 4. 225–236.

Lambrix P., Tan H. Ontology Alignment and Merging // Ontologies for Bioinformatics Principles and Practice. 2008. 6. 133–149.

Lambrix P., Tan H., Jakonienè V., Ströombäck L. Biological Ontologies // Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. 2007. 4. 85–99.

Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals // Soviet Physics Doklady. 1966. 10, 8. 707–710.

Li F., Liu H., Zhang D., Ma Y., Zhu B. Metabolic plasticity and regulation of T cell exhaustion // Immunology. 2022. 167, 4. 482–494.

Loison L. The environment: An ambiguous concept in Waddington's Biology // Studies in History and Philosophy of Science. 2022. 91. 181–190.

Lord Tessa, Nixon Brett. Metabolic changes accompanying spermatogonial stem cell differentiation // Developmental Cell. 2020. 52, 4. 399–411.

Mabbott N. A, Baillie J., Brown H., Freeman T. C., Hume D. A. An expression atlas of human primary cells: Inference of gene function from Coexpression Networks // BMC Genomics. 2013. 14, 1. 632.

Maherali N., Sridharan R., Xie W., Utikal J., Eminli S., Arnold K., Stadtfeld M., Yachechko R., Tchieu J., Jaenisch R., Plath K., Hochedlinger K. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution // Cell Stem Cell. 2007. 1, 1. 55–70.

Malone J., Holloway E., Adamusiak T., Kapushesky M., Zheng J., Kolesnikov N., Zhukova A., Brazma A., Parkinson H. Modeling sample variables with an experimental factor ontology // Bioinformatics. 2010. 26, 8. 1112–1118.

Merrell A. J., Stanger B. Z. Adult cell plasticity in vivo: De-differentiation and transdifferentiation are back in style // Nature Reviews Molecular Cell Biology. 2016. 17, 7. 413–425.

Mi H., Muruganujan A., Thomas P. D. Panther in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees // Nucleic Acids Research. 2012. 41, D1.

Mootha V. K., Lindgren C. M., Eriksson K. F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstråle M., Laurila E., Houstis N., Daly M., Patterson N., Mesirov J. P., Golub T. R., Tamayo P., Spiegelman B., Lander E. S., Hirschhorn J. N., Altshuler D., Groop L. C. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes // Nature Genetics. 2003. 34, 3. 267–273.

Mukherjee S. B., Detroja R., Mukherjee S., Frenkel-Morgenstern M. The landscape of expressed chimeric transcripts in the blood of severe COVID-19 infected patients // Viruses. 2023. 15, 2. 433.

Muralidharan S., Mandrekar P. Cellular stress response and innate immune signaling: Integrating pathways in host defense and inflammation // Journal of Leukocyte Biology. 2013. 94, 6. 1167–1184.

Müller-Molina A. J., Schöler H. R., Araúzo-Bravo M. J. Comprehensive Human Transcription Factor Binding Site Map for Combinatory Binding Motifs Discovery // PLoS ONE. 2012. 7, 11. 1–13.

Noble D. Conrad Waddington and the origin of Epigenetics // Journal of Experimental Biology. 2015. 218, 6. 816–818.

Nyamathulla S., Ratnababu P., Shaik Nazma S., Lakshmi B. A Review on Selenium Web Driver with Python // Annals of R.S.C.B. 2021. 25, 4. 16760–16768.

Ono S. J., Liou H. C., Davidon R., Strominger J. L., Glimcher L. H. Human X-box-binding protein 1 is required for the transcription of a subset of human class II major histocompatibility genes and forms a heterodimer with c-fos. // Proceedings of the National Academy of Sciences. 1991. 88, 10. 4309–4312.

Osumi-Sutherland D. Cell ontology in an age of data-driven cell classification // BMC Bioinformatics. 2017. 18, 17. 558.

Papalexi E., Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity // Nature Reviews Immunology. 2017. 18, 1. 35–45.

Rajagopal J., Stanger B. Plasticity in the adult: How should the waddington diagram be applied to regenerating tissues? // Developmental Cell. 2016. 36, 2. 133–137.

Rapin N., Bagger F. O., Jendholm J., Mora-Jensen H., Krogh A., Kohlmann A., Thiede C., Borregaard N., Bullinger L., Winther O., Theilgaard-Monch K., Porse B. Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients // Blood. 2014. 123, 6. 894–904.

Riva C., Hajduskova M., Gally C., Suman Shashi K., Ahier A., Jarriault S. A natural transdifferentiation event involving mitosis is empowered by integrating signaling inputs with conserved plasticity factors // Cell Reports. 2022. 40, 12. 111365.

Rozenblatt-Rosen O., Stubbington M. J. T., Regev A., Teichmann S. A. The Human Cell Atlas: from vision to reality. // Nature. 2017. 550. 451–453.

Sarntivijai S., Lin Y., Xiang Z., Meehan T. F., Diehl A. D., Vempati U. D., Schürer S. C., Pang C., Malone J., Parkinson H., Liu Y., Takatsuki T., Saijo K., Masuya H., Nakamura Y., Brush M. H., Haendel M. A., Zheng J., Stoeckert C. J., Peters B., Mungall C. J., Carey D. J., Athey B. D., He Y. CLO: The Cell Line Ontology // Journal of Biomedical Semantics. 2014. 5, 1. 37.

Sas A. R., Carbajal K. S., Jerome A. D., Menon R., Yoon C., Kalinski A. L., Giger R. J., Segal B. M. A new neutrophil subset promotes CNS neuron survival and axon regeneration // Nature Immunology. 2020. 21, 12. 1496–1505.

Sehgal M. S., Gondal I., Dooley L., Coppel R. Coalesce gene regulatory network reconstruction: A cross-platform transcriptional gene network fusion framework // TENCON 2006 - 2006 IEEE Region 10 Conference. 2006. 10.

Sekiya S., Suzuki A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors // Nature. 2011. 475, 7356. 390–393.

Seltmann S., Stachelscheid H., Damaschun A., Jansen L., Lekschas F., Fontaine J.F., Nguyen-Dobinsky T.N., Leser U., Kurtz A. CELDA - an ontology for the comprehensive representation of cells in complex systems // BMC Bioinformatics. 2013. 14. 228.

Shen C. N., Horb M. E., Slack J. M. W., Tosh D. Transdifferentiation of pancreas to liver // Mechanisms of Development. 2003. 120, 1. 107–116.

Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L. J., Eilbeck K., Ireland A., Mungall C. J., Leontis N., Rocca-Serra P., Ruttenberg A., Sanson S. A., Scheuermann R. H., Sha N., Whetzel P. L., Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration // Nature Biotechnology. 2007. 25, 11. 1251–1255.

Son E. Y., Ichida J. K., Wainger B. J., Toma J. S., Rafuse V. F., Woolf C. J., Eggan K. Conversion of mouse and human fibroblasts into functional spinal motor neurons // Cell Stem Cell. 2011. 9, 3. 205–218.

Subramanian A., Tamayo P., Mootha V. K., Mukherjee S., Ebert B. L., Gillette M. A., Paulovich A., Pomeroy S. L., Golub T. R., Lander E. S., Mesirov J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles // Proceedings of the National Academy of Sciences. 2005. 102, 43. 15545–15550.

Szabo E., Rampalli S., Risueño R. M., Schnerch A., Mitchell R., Fiebig-Comyn A., Levadoux-Martin M., Bhatia M. Direct conversion of human fibroblasts to multilineage blood progenitors // Nature. 2010. 468, 7323. 521–526.

Sáez M., Briscoe J., Rand D. A. Dynamical landscapes of Cell Fate Decisions // Interface Focus. 2022. 12, 4.

Sánchez-Romero M. A., Casadesús J. Waddington's landscapes in the bacterial world // Frontiers in Microbiology. 2021. 12.

Takahashi K. Cellular reprogramming – lowering gravity on Waddington's epigenetic landscape // Journal of Cell Science. 2012.

Takahashi K., Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors // Cell. 2006. 126, 4. 663–676.

Thellin O., Zorzi W., Lakaye B., De Borman B., Coumans B., Hennen G., Grisar T., Igout A., Heinen E. Housekeeping genes as internal standards: Use and limits // Journal of Biotechnology. 1999. 75, 2–3. 291–295.

Tilli T. M., Castro C. da, Tuszynski J. A., Carels N. A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases // BMC Genomics. 2016. 17, 1.

Trapnell C. Defining cell types and states with single-cell genomics // Genome Research. 2015. 25, 10. 1491–1498.

Tukey J. W. Exploratory Data Analysis. 1977. 714–719.

Tyurin-Kuzmin P. A., Molchanov A. Y., Chechekhin V. I., Ivanova A. M., Kulebyakin K. Y. Metabolic regulation of mammalian stem cell differentiation // Biochemistry (Moscow). 2020. 85, 3. 264–278.

Waddington C. H. Canalization of development and the inheritance of acquired characters // Nature. 1942. 150, 3811. 563–565.

Waddington C. H. The strategy of the genes. 1957.

Waddington C. H. Evolutionary adaptation // Perspectives in Biology and Medicine. 1959. 2, 4. 379–401.

Wagner F. pyaffy: An efficient python/cython implementation of the RMA method for processing raw data from Affymetrix Expression microarrays // PeerJPreprints. 2016.

Wang G., Heijs B., Kostidis S., Rietjens R. G. J., Koning M., Yuan L., Tiemeier G. L., Mahfouz A., Dumas S. J., Giera M., al. et. Spatial dynamic metabolomics identifies metabolic cell fate trajectories in human kidney differentiation // Cell Stem Cell. 2022. 29, 11.

Whitley D. A genetic algorithm tutorial // Statistics and Computing. 1994. 4, 2.

Wu F., R. Su, Y. Lai, X. Wang. Engineering of a synthetic quadrastable gene network to approach Waddington landscape and cell fate determination // eLife. 2017. 6. e23702.

Xie H., Ye M., Feng R., Graf T. Stepwise reprogramming of B cells into macrophages // Cell. 2004. 117, 5. 663–676.

Zappia L., Phipson B., Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scrna-tools database // PLOS Computational Biology. 2018. 14, 6.

Zhao B. Web scraping // Encyclopedia of Big Data. 2017. 1–3.

Agradecimientos

Todas las personas con las que he cruzado han cumplido su parte y esto es al final el fruto también de ese apoyo, por lo que tengo que agradecer a mucha gente por ser el soporte emocional y social detrás de todo esto.

Marcos y Daniela, gracias por aceptarme en el laboratorio. Alex, gracias por estar ahí siempre listo para ayudarme, he aprendido mucho de ti. Mikel, sin tu ayuda en el laboratorio me habría costado mucho más. Koldo, gracias por echarme un cable siempre que lo he necesitado, te estoy enteramente agradecido. Itziar Frades, gracias por tus consejos y compartir tu experiencia vital conmigo.

Olga, muchas gracias por estar ahí siempre, te has convertido en mi hermana, y aprovecho para agradecer también a Urki y Sahats tus gatas.

Jesús y Luisa sois mi segunda familia, la de Donostia, muchas gracias por acogerme así.

Maitane, Maialen, Carmen, Amaia y Ainara, de verdad chicas, gracias por acogerme así, gracias a vosotras siento que esto es un poco más mi hogar.

Isipi, gracias por estar ahí siempre. Me alegro mucho de haberte encontrado en medio de todo este torbellino.

Iñaki, mi ex-pisukide. Pasamos una pandemia. Me alegro mucho de que coincidiéramos y en mí tienes un colega para toda la vida.

Ana, gracias por ayudarme a entenderme y a lidiar mejor con mis problemas, sin ti no habría podido terminar.

Jose y Lore, mil gracias por estar pendientes de mí cuando voy a Barbastro y recibirme siempre. Siempre habéis estado ahí los dos (solo que Jose un poquito antes jeje) y os doy las gracias por ello.

George, gracias por ayudarme a echar unas risas y a desconectar cuando hace falta. Ivan, desde siempre has sido un estímulo intelectual y me alegro de que nos hayamos cruzado en el camino. También quiero agradecer a mi grupo de rol (Lizer, Ivan, Iris, Demar, Sofía y Hector) gracias por estar ahí a través de internet (y en persona para tomar algo) y amenizarme muchas tardes, espero que continúen.

Vadillo, me alegro mucho de que hayas vuelto a mi vida y espero que sigas en ella. Pablo gracias por tu apoyo, tío. Te has convertido en uno de mis mejores amigos. Sara, gracias por estar ahí, sobretodo los primeros años, sé que no fue fácil, pero sin tu apoyo tampoco sé si habría terminado.

Y en general agradecer también a la Peña Pómez, gracias por dejarme ser vuestro invitado, esas noches me han ayudado enormemente a desconectar.

También me gustaría agradecer a mi familia, tanto a la de Berbegal como a la de Barbastro por el apoyo que me habéis dado este tiempo, de verdad, mil gracias.

Me gustaría agradecer a mis padres: Mamá, Papá gracias por estar ahí y tener paciencia conmigo, ha sido un camino de altibajos (lo sabéis) pero por fin tengo la tesis. Espero pasar mucho más tiempo con vosotros ahora.

No quiero dejar de dar las gracias a mi abuela Benita, mi segunda madre. No estás aquí, yaya, pero me siento enormemente afortunado por el tiempo que compartí contigo.