Centro de Investigación en Matemáticas, A.C.

CIMAT

Modelo estadístico para la asignación de ARNs pequeños, en el contexto de interacción huésped-parásito

T E S I S

Que para obtener el grado de
Maestro en Ciencias
con especialidad en
Probabilidad y Estadística

P r e s e n t a:
Gilberto Flores Vargas

Director de tesis:
Miguel Nakamura Savoy

Autorización de la versión final

Guanajuato, Gto. Agosto de 2019

Centro de Investigación en Matemáticas, A.C.

Acta de Examen de Grado

Acta No.: 149

Libro No.: 002

Foja No.: 149

En la Ciudad de Guanajuato, Gto., siendo las 16:00 horas del día 29 de agosto del año 2019, se reunieron los miembros del jurado integrado por los señores:

DR. ROLANDO JOSÉ BISCAY LIRIO (CIMAT)
DR. ROGELIO RAMOS QUIROGA (CIMAT)
DR. MIGUEL NAKAMURA SAVOY (CIMAT)

bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

GILBERTO FLORES VARGAS

en cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"MODELO ESTADÍSTICO PARA LA ASIGNACIÓN DE ARNS PEQUEÑOS, EN EL CONTEXTO DE INTERACCIÓN HUÉSPED-PARÁSITO"

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a):

APROBADO

DR. ROLANDO JOSÉ BISCAY LIRIO
Presidente

DR. ROGELIO RAMOS QUIROGA
Secretario

DR. MIGUEL NAKAMURA SAVOY
Vocal

CIMAT
DIRECCIÓN GENERAL

DR. VÍCTOR MANUEL RIVERO MERCADO
Director General

Dedico esta tesis a mis padres y a aquellos a quienes les resulte de utilidad.

Agradecimientos

A mis padres. Mi padre, Francisco Flores Domínguez, me ha enseñado desde muy pequeño que los problemas son para abordarse, para resolverlos y, en caso de que no sea posible, al menos comprenderlos. Mi madre, Catalina Vargas Sánchez, me ha motivado constantemente y ha sido un gusto enorme tener conversaciones con tan ávida lectora.

A mis hermanos, quienes han vivido por mí varias vidas. De ellos he aprendido bastante y han sido de las primeras personas que he admirado.

A mi asesor de tesis, Miguel Nakamura Savoy, quien ha dedicado su tiempo y sabios consejos a mi guía. De él he aprendido bastante, no sólo en el área de estadística sino en general. A él le debo varias de los conceptos más interesantes que he llegado a conocer.

Al Dr. Cei Abreu Goodger, quien me platicó el problema que fue abordado en la teis y mostró siempre una disposición enorme para resolver las múltiples dudas que nos surgieron. Por supuesto, también agradezco al Dr. Obed Ramírez Sánchez por auxiliarnos en el desarrollo de esta tesis y por su valiosísima aportación en proporcionarnos los datos preprocesados.

A mis profesores de la maestría. En especial al Dr. Rogelio Ramos Quiroga y al Dr. Rolando José Biscay Lirio quienes han sido mis sinodales.

Al Dr. Víctor Manuel Pérez Abreu Carrión, quien en los últimos años no sólo se ha convertido en uno de mis principales guías académicos sino también en un gran amigo.

A mis amigos Rolando Rojo Rodriguez, Yaír Adán Hernández Esparza, Jesús Joaquin Rojas, Mariana Hernández Luna y Judith Tavarez Rodríguez. Durante distintas etapas de mi vida me han apoyado, sobre todo en tiempos difíciles.

Al Centro de Investigación en Matemáticas (CIMAT) por todo lo que me brindó en los pasados 7 años (en Licenciatura y Maestría). Desde el personal y su amabilidad, hasta el apoyo económico. Al Consejo Nacional de Ciencia Y Tecnología (CONACYT) por otorgarme la beca de maestría durante los pasados dos años.

Finalmente, a todas las personas que han interactuado conmigo durante mi formación. Cada uno está presente en mi actuar diario.

Resumen

El presente proyecto ha sido motivado por la colaboración con el grupo de Genómica Computacional del ARN del Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO). El problema biológico de fondo abordado es un problema de frontera en dicho contexto y hasta el momento no existe un modelo probabilístico con el enfoque del desarrollado en esta tesis. Específicamente, en la tesis se postula un modelo probabilístico *ad hoc* que permite resolver el problema de asignación de lecturas de secuencias cortas de ARN, en el contexto de interacción entre un huésped y un parásito. Dicho problema consiste en discernir la procedencia específica de las lecturas entre varios posibles precursores. Por su parte, la solución se presenta en términos formales de probabilidad y estadística. Respecto al modelo de probabilidad desarrollado, éste aporta una nueva metodología, y está basado en modelos de abundancia de especies, algo que se ha considerado poco en el contexto de genética. Para plantearlo se hace uso de la terminología existente en el contexto de modelos gráficos de probabilidad. Destaca que el número de parámetros considerados es menor al de modelos previos relacionados. Así mismo, se desarrolla un esquema que permite resolver el problema implícito de clasificación. El esquema se basa en la aplicación de una variante del algoritmo EM: versión estocástica de EM.

Palabras Clave

LANGEBIO, ARN's pequeños, Modelación, Clasificación, Modelos de abundancia.

Índice

Agradecimientos III

Resumen V

1. Introducción 1

2. Antecedentes de probabilidad y estadística 5

2.1. Problema biológico 5

2.2. Modelo en Li *et al.* 7

2.3. Artículo Salzman *et al.* 9

2.4. Aproximación heurística 10

3. Modelo de probabilidad 13

3.1. Necesidad de un modelo probabilístico 14

3.2. Modelo abstracto 15

3.3. Análisis exploratorio de datos 16

3.4. Modelo concreto 20

3.4.1. Distribuciones condicionales específicas 21

4. Ajuste y simulación para el modelo propuesto **27**

4.1. Estimación de parámetros del modelo 28

4.1.1. Estimación por máxima verosimilitud 28

4.1.2. Estimación por mínima distancia 30

4.1.3. Estimación empleando norma $|| \cdot ||_{L_1}$ 32

4.2. Simulación de datos 33

5. Aportes del modelo: solución a problema de clasificación y discusión. **37**

5.1. Propuesta de solución vía variante estocástica del algoritmo EM 37

5.1.1. Aplicación de propuesta en datos en Bermudez-Barrientos et al. (2019) 38

5.1.2. Nota sobre implementación de algoritmo EM en este contexto 41

5.1.3. Resultados obtenidos para datos en Bermudez-Barrientos et al. (2019) 42

5.2. Comentarios finales 44

Apéndice A. Herramientas de probabilidad y estadística. **47**

A.1. Modelos gráficos dirigidos 47

A.2. Modelos de abundancia de especies 48

A.3. Algoritmo EM 49

Referencias **51**

Índice de figuras

2.1. Modelo gráfico en Li et al. (2009). 8

3.1. Modelo propuesto para la dependencia de variables. 15
3.2. Histograma del número de repeticiones de lecturas. 17
3.3. Gráfica de línea del número de repeticiones de lecturas de contig 1. 18
3.4. Histograma del número de repeticiones de lecturas de contig 1 después de modificación en lectura ambigua. 19
3.5. Asociación de lecturas por contig. 20
3.6. Ejemplos de distribución de lecturas en contigs. 21
3.7. Proporción de lecturas por individuo asociadas a cada contig. 22
3.8. Ajuste de modelo para contig dado el individuo en escala log. 23
3.9. Ejemplos de distribución de lecturas por contigs con ajuste. 24

4.1. Ajuste del modelo en escala log. 31
4.2. Muestra de ajuste en distintos contigs. 32
4.3. Distribución de los parámetros estimados. 33
4.4. Comparación del comportamiento de los distintos modelos que se pueden considerar para el paso 3. 34

4.5. Comparación entre datos observados y simulados. 35

5.1. Valor del parámetro estimado para la distribución del individuo en distintas iteraciones. 43

5.2. Ejemplos de parámetros estimados en distintas iteraciones. 44

Índice de tablas

3.1. Cuantiles empíricos del número de repeticiones de lecturas. 17

4.1. Estimaciones por individuo. 31

5.1. Lecturas únicas por tipo de ambigüedad. 39
5.2. Probabilidad condicional estimada para los contigs. 40
5.3. Probabilidad condicional estimada para la lectura. 40
5.4. Probabilidad condicional de contig dada la lectura. 40
5.5. Estimaciones por individuo al aplicar EM. 43

CAPÍTULO 1

Introducción

Esta tesis tiene su génesis en un problema motivado por la asignación de lecturas de secuencias cortas de ARN en el contexto de interacción entre un huésped y un parásito. Por asignación se entiende discernir la procedencia específica de las lecturas, de entre varios posibles precursores presentes en ambos organismos. En la tesis se desarrolló un planteamiento y solución en términos formales de probabilidad y estadística del problema mencionado. La solución estadística formal fue concebida en el contexto de un problema de clasificación, y para la formulación probabilística se tomaron en cuenta aspectos propios de un experimento específico para obtener datos bioinformáticos.

El conocimiento de la situación específica que da origen al problema desarrollado se debe al acercamiento con el Dr. Cei Abreu Goodger —Investigador del grupo de Genómica Computacional del ARN del Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO)— y es presentada en Bermudez-Barrientos et al. (2019). En este artículo se reporta y discute una serie de dificultades existentes al analizar bases de datos de lecturas de de fragmentos cortos de ARN. La lecturas que se analizaron fueron obtenidas de expe-

rimentos realizados con el objetivo de estudiar la manera en que dos individuos interactúan por medio de fragmentos cortos de ARN. La principal dificultad es que, debido a distintos factores, existen algunas lecturas para las cuales es posible asociar distintas procedencias. La asignación de las repeticiones de lecturas para las cuales existe ambigüedad en cuanto a su origen es lo que se ha llamado el *problema de asignación*. En Bermudez-Barrientos et al. (2019) se proveen algunas propuestas heurísticas de solución. Sin embargo, estas carecen de una formulación estadística clara y lo anterior se refleja en aspectos que no son tomados en cuenta. Es ahí donde surge el problema y la necesidad de resolverlo.

La relevancia de dar una solución a esta situación es doble: por una parte —en el contexto de genética— permite contribuir al avance del estudio de interacción de individuos por medio de estos mecanismos y, por otra parte —en el contexto estadístico— expone algunos de los retos existentes al lidiar con datos complejos: estimación no convencional de parámetros y necesidad de modelar. Los datos con los que se desarrollará la solución son los derivados de un experimento realizado por los autores del artículo y han sido proporcionados por el Dr. Obed Ramírez Sánchez.

Para resolver el problema planteado, ha sido necesario desarrollar labores de modelación probabilística y estadística. De manera concreta, se recurrió a modelos gráficos, modelos de abundancia y una variante del algoritmo EM. La línea seguida es la que se describe a continuación: Puesto que éste puede ser abstraído como un problema de clasificación, se optó por darle solución vía un enfoque clásico consistente en la estimación de cierta probabilidad condicional de un modelo probabilístico. Dado que anteriormente no existía un modelo para una situación como la abordada, fue necesario buscar la adaptación o propuesta de alguno. Por su familiaridad en la disciplina de genética, el modelo se planteó empleando parte del lenguaje de los modelos gráficos de probabilidad. Posteriormente, para establecer un modelo concreto se observaron las características de los datos y el formato en que éstos se encontraban. Con miras de simplificar y a la vez dar una respuesta útil, se decidió proponer distribuciones basadas en modelos de abundancia de especies. Con los elementos descritos anteriormente, se desarrolló un esquema de clasificación valiéndose de una variante estocástica del algoritmo

EM presentada en Celeux & Govaert (1992).

Los principales aportes de la tesis son la formulación del problema en términos formales de probabilidad y estadística, y la propuesta de una solución vía una variante del algoritmo EM. Esto incluye la propuesta de un modelo probabilístico concreto para representar el fenómeno. Al realizar tareas de estimación para la aplicación del algoritmo EM, se propuso una forma de estimación que aparenta ser conveniente y plantea retos estadísticos a considerar en futuros estudios. Finalmente, otro aporte de la tesis es la introducción de los modelos de abundancia para la especificación de ciertas distribuciones condicionales. Cabe mencionar que los modelos de abundancia son muy comunes en ecología pero no así en el área de genética. Estos modelos permiten reducir las tareas de inferencia sobre múltiples parámetros a tareas de inferencia sobre un número reducido de parámetros. El planteamiento anterior proporciona una nueva posibilidad para el análisis de datos en genética.

El desarrollo posterior del documento se divide en 4 capítulos. En el Capítulo 2 se discute la literatura existente relacionada al problema que nos concierne y comienza con la explicación del problema a grosso modo en términos biológicos, en el Capítulo 3 se desarrolla un modelo *ad hoc* para la situación específica, en el Capítulo 4 se simulan datos sintéticos y se discute el ajuste a los datos, y el Capítulo 5 está destinado a las conclusiones finales, la solución al problema de clasificación y alcances del modelo propuesto. Finalmente, aspectos técnicos y generales se presentan en el Apéndice A.

CAPÍTULO 2

Antecedentes de probabilidad y estadística

En el presente capítulo se comienza con una descripción precisa del problema en términos biológicos y se presentan de manera coloquial los objetos que serán sujetos de interés en los capítulos siguientes. Se continúa con una discusión de la literatura existente relacionada con el problema que se aborda. Finalmente, se comenta una propuesta heurística de Bermudez-Barrientos et al. (2019) que se formaliza y enriquece en el presente proyecto.

2.1. Problema biológico

Toda célula de cada organismo vivo posee su propio genoma —una larga secuencia de elementos que contienen código genético—. Dicho genoma contiene ciertas regiones que pueden ser transcritas mediante procesos biológicos; esto es, regiones que por medio de un proceso de transcripción de ciertos fragmentos de éstas dan origen a fragmentos de ARN. Dependiendo del material genético analizado, estas regiones pueden ser llamadas genes o de otra manera. En nuestro caso, por simplicidad, llamaremos a cada una de estas regiones un **contig**. Ahora bien, de cada contig, se puede generar cierta variedad de fragmentos de ARN

que llamaremos **isoformas**. Los fragmentos de ARN reciben distintos nombres dependiendo su longitud o función. Los fragmentos de ARN a los que se estará haciendo referencia serán los llamados fragmentos cortos de ARN[1]. La producción de isoformas varía en intensidad de contig a contig y a esta intensidad suele conocerse como nivel de expresión. Uno de los problemas principales en genética es estimar el nivel de expresión de estas regiones. Ahora bien, para realizar lo anterior, lo común es que las isoformas o fragmentos de ARN sean procesadas mediante un aparato tecnológico llamado secuenciador y traducidas a lecturas. Formalmente, las lecturas son palabras formadas con elementos del alfabeto $\{A, T, G, C\}$. Posteriormente, puede aplicarse una serie de técnicas propias de bioinformática para —partiendo de estas lecturas— inferir los niveles de expresión de interés.

En Bermudez-Barrientos et al. (2019) se estudian retos existentes al manipular datos de lecturas de ARN provenientes de ciertos experimentos específicos. Éstos experimentos buscan captar aspectos de la interacción entre células via fragmentos cortos de ARN. Uno de tales experimentos, desarollado en ese artículo, se describe a continuación. De las células de un parásito se extrae material genético y éste se inyecta a una célula de un organismo huésped. Posteriormente, en diferentes tiempos se realizan lecturas de fragmentos cortos de ARN y éstas son estudiadas. Algo que se estará realizando en lo posterior es considerar un tiempo fijo en el que ha sido analizada la célula. Por medio de técnicas de bioinformática se reconstruyen los contigs de los cuales provienen y a su vez estos son *alineados* —mediante un proceso llamado **alineación** que determina la posición en el genoma de la cual provienen secuencias de ARN— al genoma del huésped y del parásito. Con esta información es posible inferir el contig e individuo de procedencia de cada lectura, aunque no de manera unívoca; existen lecturas que pueden ser asociadas a múltiples contigs o contigs que pertenecen tanto al huésped como al parásito.

Esta incertidumbre sobre la procedencia de las lecturas surge debido a la combinatoria de los objetos de interés y las técnicas empleadas para llevar a cabo la *alineación*. A las lecturas con posible procedencia en más de un contig o individuo se les conoce como **lecturas**

[1]Fragmentos formados por 18 a 30 nucléotidos.

ambiguas. Entonces, la problemática que surge es, ¿cómo tratar las lecturas ambiguas de tal manera que puedan ser asignadas y esto permita realizar tareas de estimación del nivel de expresión de cada contig? La pregunta anterior se complica en este caso pues al tratarse de un contexto en el que se consideran dos individuos, la ambigüedad puede provenir de contigs que son compartidos tanto por uno como por otro.

En este punto es donde surge la inquietud y el acercamiento a técnicas y formulación de este problema en términos precisos de probabilidad y estadística. El problema planteado anteriormente no se ha tratado como tal en la literatura existente. Sin embargo, existen dos fuentes específicas en las que se ha basado el presente proyecto y se han empleado como un medio de introducción a los problemas específicos de genética: Salzman et al. (2011) y Li et al. (2009). Estos tratan el problema de ambigüedad o solo una parte de éste para el caso de un solo individuo y no aprovechan las ventajas conceptuales que se ganan cuando se plantea como un problema de clasificación.

2.2. Modelo en Li *et al.*

El modelo desarrollado en Li et al. (2009) es el más cercano —en cuanto al enfoque— al modelo propuesto. El objetivo de ese artículo es proponer una técnica para estimar la expresión genética aún con incertidumbre en el *alineamiento*. Se decidió estudiar dicho artículo por su relación clara con el problema que se trata en esta tesis. En su desarrollo se consideran cuatro variables aleatorias; la isoforma (G), el orden en que son leídos los fragmentos de ARN (O), posición de inicio de lectura (S) y la lectura misma (R). En su caso se consideran como no observados las variables de isoforma, posición, y orientación. Por otra parte, en este artículo suponen un catálogo de isoformas conocido y la parte aleatoria que influye de mayor manera a la ambigüedad en la asignación se encuentra en el error de medición del dispositivo de lectura.

A continuación se presenta con más detalle el modelo específico de este artículo. El modelo gráfico dirigido(ver Sección A.1 del Apéndice A) es el que se presenta en la Figura 2.1.

En este caso la distribución conjunta está dada por la relación

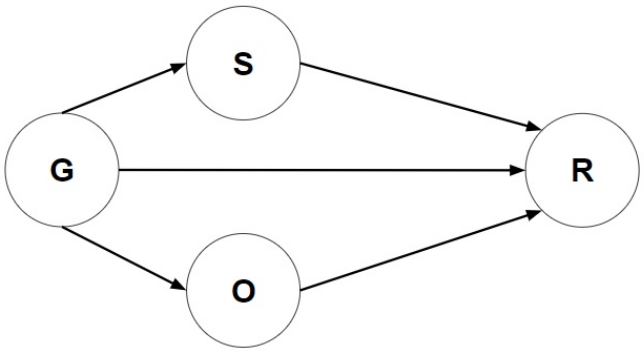

Figura 2.1: Modelo gráfico en Li et al. (2009). En Li et al. (2009) el objetivo es realizar estimaciones de expresión genética aún con problemas de ambigüedad. En su caso, el componente aleatorio al que se le da mayor énfasis es al derivado del aparato de medición. G denota la isoforma, S la posición de inicio, O el orden de lectura, y R la lectura.

$$\begin{aligned}&\mathbb{P}\left(G=g,S=s,O=o,R=r\right)\\&=\mathbb{P}\left(G=g\right)\mathbb{P}\left(S=s|G=g\right)\mathbb{P}\left(O=o|G=g\right)\mathbb{P}\left(R=r|S=s,O=o,G=g\right).\end{aligned}\tag{2.1}$$

En general se puede suponer O fija pues ésta depende del protocolo de medición y es controlable. Una observación pertinente respecto a la modelación desarrollada en Li et al. (2009) es que para cada isoforma g existe un parámetro en el modelo. Específicamente, se supone que $\mathbb{P}(G=g)=\theta_g$ y en principio no existe una relación funcional entre los θ_g salvo que $\sum_{g\in\mathcal{C}}\theta_g=1$ y $\theta_g\geq 0$, donde $\mathcal{C}$ es el catálogo de isoformas. Además, se supone que la distribución de S y $R|S=s,O=o,G=g$ son conocidas, especificadas por el investigador. Es de notar que formalmente estas serían distribuciones a estimar. En el artículo se proponen distribuciones que están basadas en estimaciones en otros experimentos. Por otra parte, respecto a la distribución condicional de las lecturas, en Li et al. (2009) se emplea la denominada *matriz de sustitución*; dicha matriz codifica las probabilidades de que un nucléotido de la isoforma g sea leído como si se tratase de otro.

En Li et al. (2009) se estiman los parámetros de interés empleando la formulación clásica del algoritmo EM desarrollada en Dempster et al. (1977) para el caso de datos no observados.

Cabe mencionar que en Li et al. (2009) no se hace explícito el problema de clasificación existente y su objetivo, como el título lo menciona, es estimar los niveles de expresión genética.

El planteamiento de Li et al. (2009) no puede ser empleado de manera literal en nuestro contexto pues no se considera el escenario de interacción entre dos individuos. Además, por la longitud pequeña de los fragmentos de ARN considerados, se puede suponer que no existe error de medición y por lo tanto no es necesario modelarlo. Las ideas retenidas de tal artículo son el uso de modelos gráficos dirigidos para la especificación de las distribuciones condicionales, y la suposición del catálogo conocido de —en nuestro caso— contigs.

2.3. Artículo Salzman *et al.*

Otro artículo que se ha consultado y analizado es Salzman et al. (2011). El objetivo de éste es estudiar el problema de estimación de niveles de expresión con estadística formal y mostrar cómo diferentes protocolos de lectura afectan la calidad de las estimaciones. Su modelo presentado es una extensión de uno previo (Jiang & Wong, 2009) basado en conteos Poisson para la distribución del número de lecturas de cada tipo. Un supuesto crucial en este artículo radica en considerar conocidos los catálogos de lecturas e isoformas o transcritos[2]. Ahora bien, lo que se supone es que la abundancia o expresión de cada trascrito i está dada por un parámetro $\theta_i > 0$. A su vez, el número de lecturas del tipo j, provenientes del transcrito i, siguen una distribución Poisson. El parámetro de ésta es $\theta_i a_{ij}$, con a_{ij} un número no negativo que puede ser interpretado como la intensidad de muestreo de la lectura j del transcrito i.

Posteriormente, en el artículo, se realiza una caracterización estadística de la distribución inducida para el número de lecturas obtenidas de cada tipo consistente en el cálculo de los estadísticos suficientes, información de Fisher, etc.

Es de destacar que en Salzman et al. (2011) se hace la aclaración específica que no se lidia con ningún problema de asignación pero que sería relevante hacerlo. De este artículo se han tomado dos ideas esenciales para el modelo que posteriormente proponemos. A saber,

[2]Estos conceptos varían un tanto pero para fines de comparación con respecto al modelo de Li et al. (2009) se considerarán equivalentes.

estas son: cada transcrito —que en nuestro caso serán los contigs— aporta cierta proporción de lecturas, y a su vez, cada lectura se manifiesta en cierta proporción de cada transcrito. Una observación pertinente es que existe una gran cantidad de parámetros a considerar. Por una parte se debe considerar un parámetro θ_i y parámetros a_{ij}, con j indexando las lecturas que pueden ser asociadas al transcrito i, para cada transcrito i.

En el modelo desarrollado en el Capítulo 3 se considera un total de uno a cuatro parámetros unidimensionales asociados por contig. Lo anterior establece una diferencia considerable con respecto a Salzman et al. (2011), donde el parámetro que resulta vincular lecturas e isoformas es de dimensión tan grande como el número de lecturas asociadas a dicha isoforma.

2.4. Aproximación heurística

Finalmente, en Bermudez-Barrientos et al. (2019) se presenta una aproximación a la solución de este problema que ya contiene algunos de los elementos formalizados en el modelo propuesto en el Capítulo 3. La idea empleada es asignar las lecturas ambiguas con base a la proporción de lecturas asociadas a cada contig. Es decir, si la lectura l_1 puede ser asociada a los contig c_1 y c_2, las repeticiones observadas de l_1 son asociadas a los contigs de la siguiente manera: si c_1 tiene asociadas n_1 lecturas y c_2 tiene asociadas n_2 lecturas, la proporción de repeticiones de la lectura l_1 asociadas al contig c_1 serán el $n_1/(n_1+n_2) \times 100\,\%$ y las asociadas al contig c_2 serán el $n_2/(n_1+n_2) \times 100\,\%$.

A pesar de que la idea descrita anteriormente es muy intuitiva y natural, carece de algunas consideraciones importantes: la abundancia de lecturas de cada individuo y la abundancia de la lectura específica en cada uno de los contigs. Dicho de manera coloquial, los términos que faltan de considerar son qué tantas lecturas aporta cada individuo y qué tantas lecturas de ese tipo aportaría cada contig. Por otra parte, es de observar que en cierta manera está considerándose la idea de asignar las repeticiones de las lecturas con base a cierta probabilidad estimada. Uno de los aportes principales de la presente tesis es la incorporación de estos términos mediante un modelo probabilístico y la formulación de este problema en términos

formales de probabilidad y estadística posibilitando así el empleo de técnicas específicamente diseñadas para resolver problemas de clasificación.

CAPÍTULO 3

Modelo de probabilidad

El presente capítulo inicia argumentando la necesidad de plantear un modelo de probabilidad con miras a ser empleado para tareas de clasificación y estimación. Posteriormente se especifican los supuestos que se realizarán para desarrollarlo. Hechos explícitos los supuestos, se desarrolla el modelo en abstracto. Con el fin de buscar propuestas concretas para las distribuciones específicas de los componentes del modelo, se realiza un análisis exploratorio de los datos. Finalmente se sintetizan las secciones anteriores por medio de la propuesta del modelo específico.

Cabe mencionar que el plantear un modelo de probabilidad surge por la necesidad de clasificar de manera formal —con el fin ulterior de realizar tareas de estimación— las lecturas ambiguas en el experimento específico descrito en el Capítulo 2. En consecuencia, varios de los aspectos considerados en el modelo tienen su génesis de manera directa en los datos, incorporando los razgos esenciales que participan en su producción.

Con fines de formular en términos formales de probabilidad y estadística el problema que nos concierne, los objetos de interés se han abstraído por medio del concepto de variable

aleatoria. Por la discusión del Capítulo 2, las variables que se estarán considerando en lo sucesivo son las siguientes tres: lectura observada L, contig asociado C, e individuo I. La característica esencial del problema en cuestión puede tratarse como un problema de clasificación donde existen realizaciones de L para las cuales no se conoce el valor específico de C ni el de I.

3.1. Necesidad de un modelo probabilístico

Uno de los resultados fundamentales en el contexto de clasificación es el del **clasificador bayesiano óptimo**. Éste establece que la mejor manera de clasificar se basa en la probabilidad condicional de la clase dado el dato observado. Por este motivo, una de las tareas primordiales de clasificación consiste en realizar una estimación de dicha probabilidad condicional. Para lo anterior es imprescindible contar con algún modelo de probabilidad.

En el contexto específico de la interacción de dos especies, aunado a que el interés radica en el análisis de secuencias cortas de ARN, un modelo genérico —como los comúmente empleados para realizar tareas de clasificación— resultaría poco útil. Por lo tanto, se requiere un modelo que reconozca e incorpore los elementos pertinentes y omita aquellos que en el contexto resultan superfluos. Otra razón para no considerar un modelo genérico es que la tarea de clasificación no constituye el fin último del proyecto de investigación en Bermudez-Barrientos et al. (2019), pues un objetivo posterior es realizar inferencia sobre los parámetros propuestos. Ahora bien, dado que el interés radica en inferir para qué clase es mayor la probabilidad condicional, la probabilidad condicional de la clase dado el dato observado puede ser sustituida por la probabilidad conjunta del dato observado y la clase. Entonces, basta con conocer la probabilidad conjunta mencionada. Por esta razón, lo que se realizará a continuación es plantear un modelo mediante el cual se especifique la probabilidad conjunta de las variables aleatorias de interés.

3.2. Modelo abstracto

De manera análoga a Salzman et al. (2011), se supone que el catálogo de lecturas y contigs es el observado y que es conocido. Además, dado que las lecturas obtenidas provienen del procesamiento de secuencias cortas de ARN, se supone que no existe error de medición. Por lo tanto, el modelo desarrollado no tendrá en consideración este aspecto. Es decir, las lecturas se pueden considerar equivalentes a las llamadas *isoformas* en Li et al. (2009), en el sentido de que los fragmentos de ARN se secuencian sin error.

Debido a la familiaridad en el contexto de bioinformática y flexibilidad que ofrecen para modelar, se ha recurrido a los modelos gráficos para especificar la distribución conjunta de las variables I, C y L. Esta está especificada por medio del modelo gráfico dirigido presentado en la Figura 3.1. Acorde al lenguaje de los modelos gráficos y la Figura 3.1, la distribución

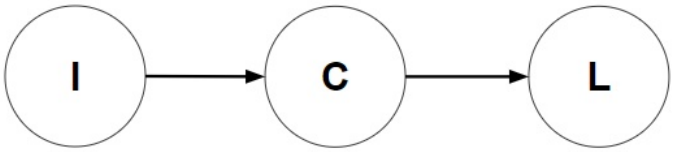

Figura 3.1: Modelo propuesto para la dependencia de variables. El modelo propuesto consta de 3 variables y sus relaciones de dependencia se especifican mediante el modelo gráfico presentado.

conjunta propuesta es, para i, c, l, la expresada en la ecuación (3.1)

$$\mathbb{P}\left(I = i, C = c, L = l\right) = \mathbb{P}(I = i)\mathbb{P}\left(C = c | I = i\right)\mathbb{P}(L = l | C = c). \tag{3.1}$$

Por simplicidad, y dado que la distribución conjunta está determinada por el producto de distribuciones condicionales, se especifica el soporte de las variables aleatorias cuando éstas han sido condicionadas a valores específicos de las variables restantes. Procediendo de esta manera, los soportes están determinados de la siguiente forma:

- el soporte de I es $\{0, 1\}$ donde 0 denota al huésped y 1 denota al parásito.
- El soporte de $C|I = i$ es $\mathcal{C}(i)$ donde $\mathcal{C}(i) = \{c_1^i, \ldots, c_{n_i}^i\}$ es el conjunto de etiquetas de los contigs del individuo i.

- Finalmente, el soporte de $L|C = c$ es $\mathcal{L}(c) = \{l_1^c, \ldots, l_{k_c}^c\}$, el catálogo de lecturas del contig c.

En principio se pueden especificar distribuciones que tomen en consideración la estructura de "palabras" de las lecturas y los contigs —algo semejante a lo realizado en Li et al. (2009) por medio de la matriz de sustitución. Sin embargo, el suponer los catálogos conocidos permite considerar familias de distribuciones conocidas en otros contextos que se adecuen a este caso. En este punto se ha optado por observar a los datos antes de proceder a postular las distribuciones específicas.

3.3. Análisis exploratorio de datos

Los datos que se describen a continuación han sido proporcionados por el Dr. Cei Abreu Goodger y corresponden a uno de los experimentos realizados en Bermudez-Barrientos et al. (2019). Gran parte del análisis ha sido posible gracias a la estructura de datos proporcionada y al uso de la paquetería `plyr` de `R`. A una escala mayor incluso el análisis presentado en esta sección se complica considerablemente y es necesario el uso de software especializado para la manipulación e integración de bases de datos.

En total se cuenta con 15589 contigs y 28671 lecturas únicas. Para cada lectura única se cuenta con el número de repeticiones de la misma. Cabe mencionar que, como lo muestran las estadísticas descriptivas presentadas en la Tabla 3.1, la distribución de estas repeticiones es muy dispar. Existen incluso 1500 lecturas que sólo cuentan con una repetición. Para facilitar el manejo de los datos estas lecturas son incluidas en lo sucesivo. Es de notar que el omitir las lecturas con una sola repetición es una decisión sujeta a las implicaciones que esta acción tenga en el estudio dentro del contexto de genética.

Es de observar que gran parte de las lecturas cuentan con cinco repeticiones o menos. Por otra parte, destaca el hecho de que existe una lectura con 696756 repeticiones. Con fines ilustrativos, en la Figura 3.2 se presenta una gráfica de barras del número de lecturas únicas contra el número de repeticiones.

Tabla 3.1: Cuantiles empíricos del número de repeticiones de lecturas únicas. Como se observa en la tabla, al menos el 50 % de las lecturas tiene cinco repeticiones o menos.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	3	5	102	8	696756

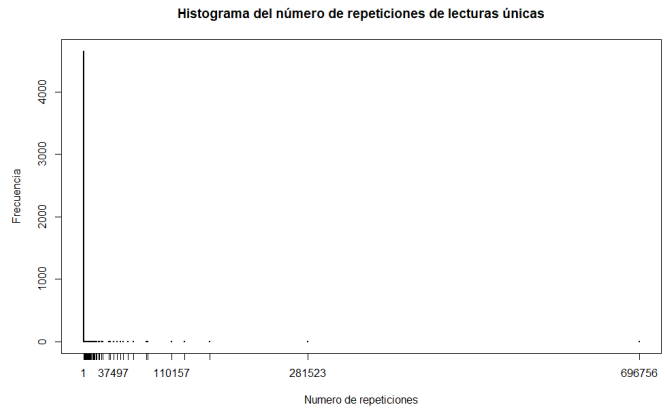

Figura 3.2: Histograma del número de repeticiones de lecturas. Graficamente se observa que son muchas las lecturas con pocas repeticiones y pocas las que tienen un gran número.

Al analizar con mayor detalle se concluye que la lectura con 696756 repeticiones se encuentra asociada al contig $c1$ cuyo origen es el huésped. Para presentar de manera intuitiva el fin perseguido al clasificar las lecturas ambiguas, se analiza con mayor detenimiento el contig en cuestión. En este contig hay asociadas dos lecturas ambiguas: una con 138 repeticiones y otra con cinco. Ordenando de mayor a menor las frecuencias de las lecturas asociadas a este contig, quitando las primeras nueve más frecuentes, se obtiene la gráfica presentada en la Figura 3.3.

Ahora bien, estudiando la figura mencionada (3.3), se puede observar que el valor de 138 parece "romper" con el patrón de decaimiento de la curva. Si esta lectura tuviera sólo 120 repeticiones asignadas al contig c1 se obtendría la gráfica de la Figura 3.4. En esta figura se observa que las 120 repeticiones ya no parecen romper con el patrón de decaimiento. A grosso modo, este tipo de adaptaciones son las pretendidas al clasificar según un modelo probabilístico las lecturas ambiguas: clasificarlas de tal manera que dicha clasificación sea

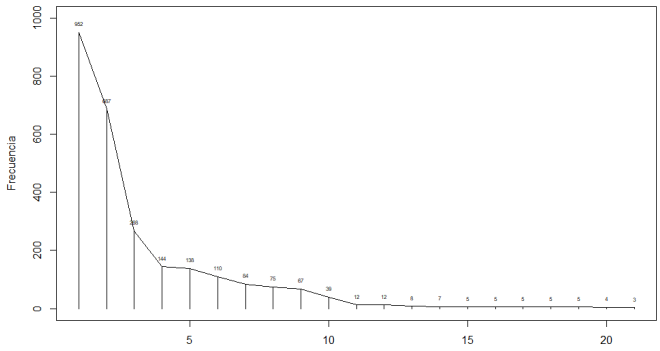

Figura 3.3: Gráfica de línea del número de repeticiones de lecturas del contig c1. Se observa un patrón claro de decaimiento tal como el presente en los modelos de abundancia de especies.

más "creíble" según el modelo propuesto.

Se continúa el análisis exploratorio de los datos observando la asociación de contigs y lecturas. Resulta ser que esta asociación es muy dispar. En la Figura 3.5 se grafica por medio de barras el número de contigs contra el número de lecturas únicas. Es decir, la coordenada horizontal corresponde al número de lecturas únicas y la coordenada vertical al número de contigs que tiene asociadas ese número de lecturas únicas. Hay contigs que sólo tienen asociada una lectura única y el número de repeticiones de esta es pequeño. Al respecto, el número de contigs que sólo tienen asociada una lectura única es 13714. Por otra parte, el contig que tiene más lecturas únicas asociadas tiene 1156.

Lo anterior insinúa que en desarrollos posteriores conviene considerar catálogos que abarquen un mayor número lecturas asociadas. En principio —-ignorando la posible relevancia biológica de estos catálogos— los elementos con pocas lecturas únicas asociadas o, viceversa, los elementos del catálogo con muchas lecturas, podrían aportar poca información.

Este fenómeno, contigs y lecturas que abarcan la mayor parte de la frecuencia observada, sugiere proponer de modelos de abundancia de especies. Con este término se hace referencia a una gama de modelos comúnmente empleados en ecología para describir la abundancia de especies en un ecosistema determinado (ver Sección A.2 en el Apéndice A).

Para realizar una exploración visual de que tan adecuada es la adopción de modelos se-

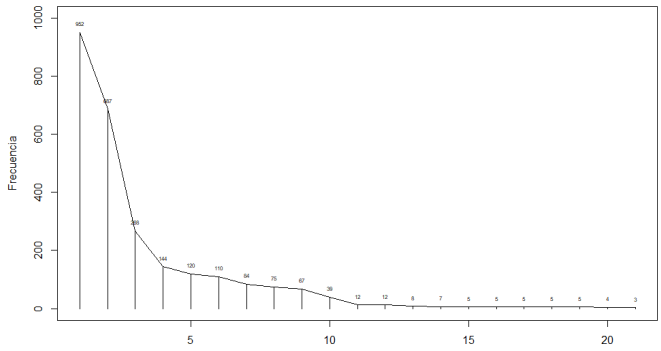

Figura 3.4: Histograma del número de repeticiones de lecturas de contig c1 después de modificación en el número de repeticiones de una lectura ambigua. Con la modificación del número de repeticiones de la lectura ambigua resulta más claro el patrón de decaimiento de la curva. Será de interés al asignar la lectura ambigua si este es el número de repeticiones que se asigna al contig en cuestión.

mejantes a los de abundancia, se han graficado los conteos de lecturas únicas asociadas a algunos contigs. Las gráficas se realizaron de la siguiente manera: Para cada contig se seleccionan las lecturas que tiene asociadas. Una vez que se identifican todas las lecturas asociadas a un contig, éstas se ordenan de mayor a menor con respecto a su número de repeticiones y posteriormente se grafican estas cantidades como se muestra en la ilustración de la Figura 3.6. En este paso, las lecturas que están asociadas a distintos contigs han sido contadas como si todas las repeticiones vinieran del mismo contig. A manera de ejemplo: si la lectura s_1 con n_1 repeticiones está asociada a los contigs c_1 y c_2, se considera como n_1 lecturas del tipo s_1 en el contig c_1 y en el contig c_2.

Como se puede observar en la Figura 3.6, son pocas las lecturas que abarcan la mayor proporción de las lecturas asociadas al contig, con un decaimiento suave y sistemático. Por lo tanto, la distribución condicional basada en modelos de abundancia de las lecturas dados los contigs parece razonable.

La adopción de modelos de abundancia en primera instancia fue motivada por datos observados. Específicamente, el concepto reinante es que algunos pocos entes acumulan la gran

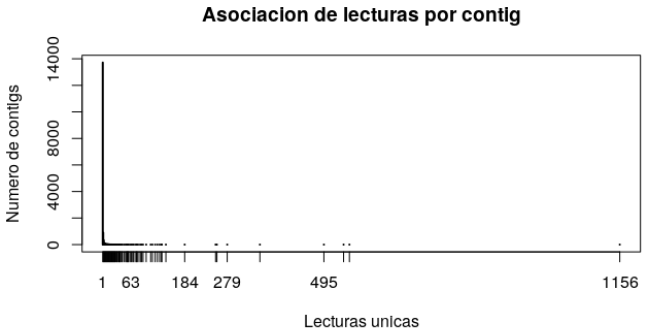

Figura 3.5: Histograma del número de lecturas asociadas por contig. En su mayoría, los contigs sólo tienen asociada una lectura única. Destaca que existe un contig con 1156 lecturas únicas asociadas.

mayoría de ocurrencias, mientras que muchos otros entes muestran manifestarse con muy poca frecuencia. En este trabajo, la forma funcional específica utilizada para describir el perfil de frecuencias observado fue propuesta empíricamente. El que dicha forma funcional posea una interpretación biológica no ha sido un asunto investigado por el momento. Por ejemplo, una interpretación como la buscada podría ser la dada por el siguiente caso: que sean pocos los fragmentos de ARN que deben generarse para que la célula funcione adecuadamente. En la literatura de modelos de abundancia, existen algunas formas funcionales que sí poseen una explicación física, pero proliferan varias instancias de modelos empíricos similares al que se ha determinado para este caso (ver sección "THEORETICAL DEVELOPMENTS IN SADS" de McGill et al. (2007)). Contar con una interpretación contribuiría a discenir cuál elegir.

3.4. Modelo concreto

Con el análisis anterior, existen las condiciones para establecer una forma específica de las distribuciones condicionales en el modelo gráfico. La distribución especificada para I es una distribución Bernoulli $\mathrm{Ber}(p)$. Partiendo del análisis llevado a cabo en la Sección 3.3, se realiza la caracterización de las distribuciones condicionales para C y L basándose en los modelos de abundancia de especies. De esta manera, la distribución de $C|I = i_0$ será

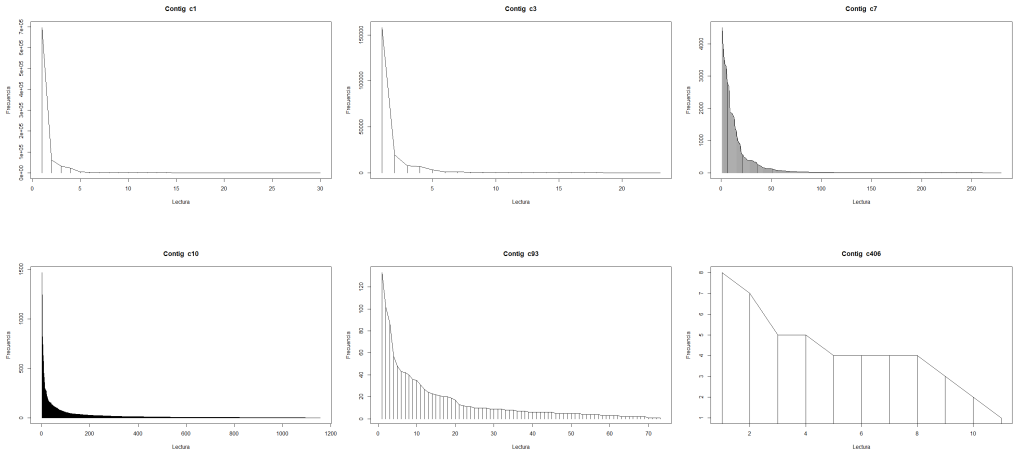

Figura 3.6: Ejemplos de la distribución del número de repeticiones de lecturas asociadas por contig. Para los contigs que tienen asociada más de una lectura se observa un patrón de decaimiento de las frecuencias característico en los modelos de abundancia.

tal que existe una permutación $\sigma_{i_0} : \{1, \ldots, n_{i_0}\} \to \{1, \ldots, n_{i_0}\}$ con la propiedad de que $\mathbb{P}\left(C = c_j^{i_0} | I = i_0\right) = f(x_{i_0}, \sigma(j))$, siendo $f : \mathbb{R} \times \mathbb{N} \to [0, 1]$ decreciente con respecto al segundo parámetro. De manera análoga se especifica la distribución condicional de $L|C = c_0$.

3.4.1. Distribuciones condicionales específicas

En el caso de la distribución del número de lecturas por contig de los asociados a cada individuo, la disparidad entre las proporciones también es muy notable. Lo anterior se ilustra con la distribución del número de lecturas por contig de los asociados al huésped presentada en la Figura 3.7.

Con el fin de explorar formas específicas para la función de probabilidad de $C|I = i$, dado lo observado en la Figura 3.7, se recurre a la escala logarítmica y se ajusta a la curva estimada con base en las frecuencias la función

$$-\alpha_{i1} \log\left(\alpha_{i2} k\right) + \alpha_{i3} \frac{k}{n_{\mathcal{C}(0)}}, \tag{3.2}$$

donde $n_{\mathcal{C}(0)}$ representa el número total de contigs asociados al huésped. Es de notar que en este caso α_{i1} debe ser mayor a α_{i3} para que la función sea decreciente con respecto a k.

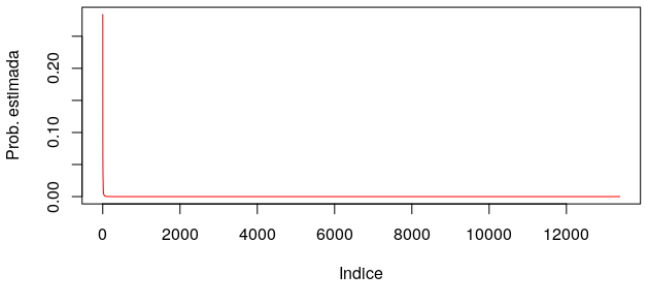

Figura 3.7: Proporción de lecturas, para el caso del huésped, asociadas a cada contig. Sobresale el contig que tiene mayor cantidad de lecturas asociadas. Para los demás contigs la gráfica es poco informativa.

Además es necesario $\alpha_{i2} > 0$.

Los resultados obtenidos al encontrar $(\alpha_{i1}, \alpha_{i2}, \alpha_{i3})$ —por medio de paquetería de `R` para optimización con restricciones, pues los parámetros deben satisfacer $\alpha_{i1} > \alpha_{i3}$ y $\alpha_2 > 0$— son los presentados en la Figura 3.8.

Si bien el ajuste de la función no es capaz de representar todos los detalles de la curva, el comportamiento global es bien reflejado. Por lo tanto, se propone la función descrita en (3.2) como el logaritmo del kernel de la distribución condicional en cuestión. Esto es, utilizando la notación introducida anteriormente, la propuesta anterior corresponde a

$$f(\alpha, k) = T^{-1} \left(\alpha_{i2} k\right)^{-\alpha_{i1}} \exp\left(\alpha_{i3} \frac{k}{n_{\mathcal{C}(0)}}\right), \tag{3.3}$$

con T una constante de normalización. Específicamente, la probabilidad condicional de los contigs dado el individuo es

$$\mathbb{P}\left(C = c_j^i | I = i\right) = T^{-1} \left(\alpha_{i2} \sigma_{(}j)\right)^{-\alpha_{i1}} \exp\left(\alpha_{i3} \frac{\sigma_i(j)}{n_{\mathcal{C}(0)}}\right). \tag{3.4}$$

Ahora se analiza el caso de la distribución condicional $L|C = c$. Debido a que el número de lecturas únicas asociadas a cada contig varía considerablemente entre contigs, se propone para esta distribución condicional un modelo más simple[1]. Empleando la notación expuesta

[1]Con fines exploratorios. Como resultado de este análisis se concluyó que conviene un modelo con mayor número de parámetros.

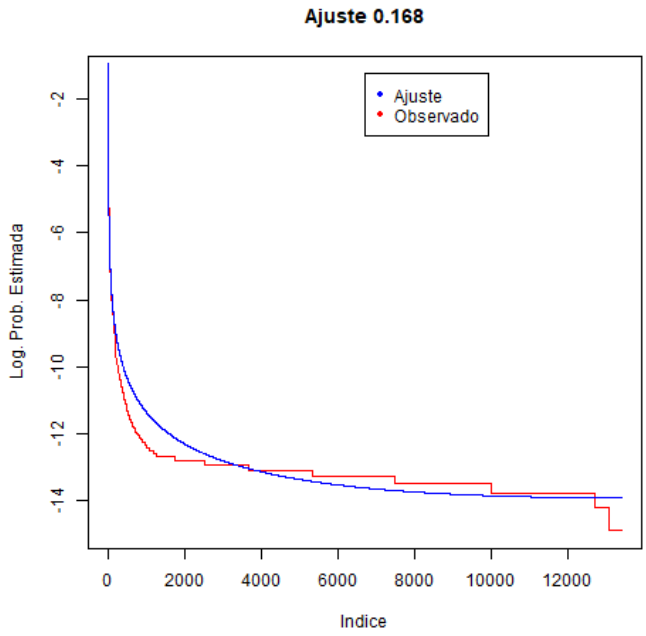

Figura 3.8: Ajuste de modelo para contig dado el individuo en escala log: caso del huésped. El valor numérico del título corresponde al valor de la función objetivo que ha sido la dada por mínimos cuadrados.

al inicio de la sección, para este caso

$$f(x,k) = r^{-1}x^k, \tag{3.5}$$

donde r es una constante de normalización y $x \in [0,1]$. Al ajustar con base a las proporciones observadas y empleando un optimizador numérico, se obtienen los resultados mostrados en la Figura 3.9. La Figura 3.9 sugiere que el kernel propuesto captura los razgos esenciales de la curva de interés. Entonces, se propone

$$\mathbb{P}\left(L = l_k | C = c\right) = r^{-1}\theta_c^{\sigma_c(k)}, \tag{3.6}$$

siendo r una constante de normalización y σ_c la permutación correspondiente.

Se observa que en general el ajuste es razonable. Sin embargo, hay casos donde las probabilidades de los elementos más abundantes están siendo subestimados. Por lo anterior, es sensato considerar modelos con más parámetros que permitan reproducir las variaciones ob-

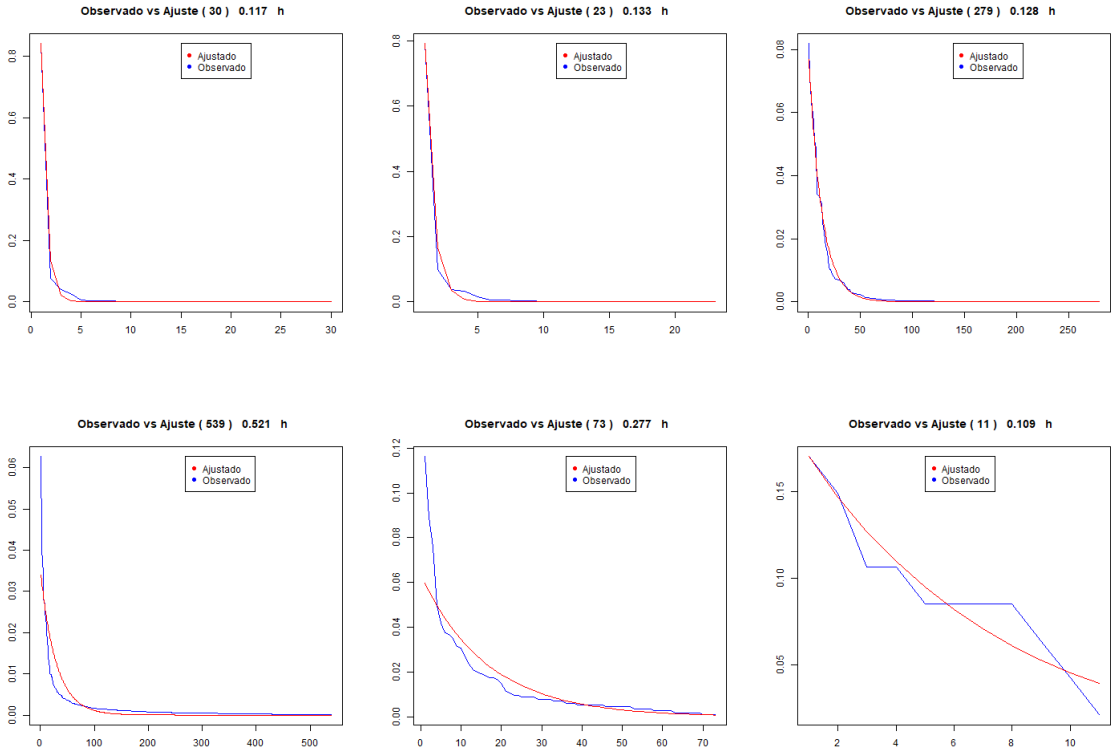

Figura 3.9: Ejemplos de distribución del número de repeticiones de lecturas por contigs con ajuste de la función propuesta. El modelo propuesto se ajusta adecuadamente. Sin embargo, difícilmente representa algunas sutilezas en la distribución y por esta razón se recomienda utilizar un modelo como el especificado para la distribución del contig dado el individuo.

servadas. Por otra parte, cabe mencionar que para los contigs que sólo tienen asociada una lectura, no es necesario considerar este modelo. Formalmente —para estos casos— al considerar el modelo propuesto, cualquier valor de θ_c es adecuado. En el siguiente capítulo se analizan con mayor detalle los resultados obtenidos por medio de este ajuste numérico.

A continuación se mencionan algunas de las diferencias con los modelos comentados en la Sección 2.2 y 2.3 del Capítulo 2. Con respecto al modelo planteado en Li et al. (2009), destaca que en el modelo desarrollado no se consideran los errores de medición. Más aún, la sucesión de letras que forman a la lectura no participa de manera directa en la distribución condicional correspondiente. A pesar de que tanto en Li et al. (2009) como en el presente

documento el modelo gráfico consta de tres nodos, estos difieren considerablemente en su interrelación. Por ejemplo, en su caso, las isoformas son un objeto análogo a los contigs pero la manera en que son tratados es distinta, ya que en ese modelo se considera un parámetro por cada isoforma. Ahora bien, con respecto al modelo de Salzman et al. (2011), no se supone una distribución específica —como la distribución Poisson— para los conteos de las lecturas sino un modelo que permite relacionar sus frecuencias con todas las demás lecturas. Al igual que en el caso de Li et al. (2009), los parámetros asociados a los objetos análogos a los contigs, son tantos como el número de éstos. Finalmente, la diferencia principal entre los modelos en Salzman et al. (2011) y Li et al. (2009) y el propuesto radica en que éste considera el escenario donde las lecturas provienen de dos individuos en lugar de sólo uno.

CAPÍTULO 4

Ajuste y simulación para el modelo propuesto

En el presente capítulo se explora cómo estimar los parámetros para el modelo propuesto —pues este será uno de los pasos más importantes al aplicar el algoritmo EM— y se comentan los resultados obtenidos al hacerlo con un juego de datos presentado en Bermudez-Barrientos et al. (2019). Dado que el fin de este capítulo es ilustrar algunas técnicas de estimación, se ignorará provisionalmente el problema de las lecturas ambiguas. Se entiende por esto que las repeticiones serán consideradas como asignadas todas sin ambigüedad. Por supuesto que la presencia de ambigüedad es el principal motivador del problema; en la Sección 5.1 se comentará sin falta la adaptación que habrá que hacer en términos de estimación en esta circunstancia, y se verá que las estimaciones aquí obtenidas no serán más que valores iniciales para cierto algoritmo iterativo. Posteriormente se simulan datos empleando el modelo propuesto.

4.1. Estimación de parámetros del modelo

En esta sección se comienza por explorar la estimación de los parámetros del modelo vía máxima verosimilitud. Para el modelo desarrollado en el Capítulo 3, con este enfoque clásico, no se obtienen resultados que se adecuen debidamente a las sutilezas de la distribución y se presentan algunos problemas prácticos al emplear métodos númericos de optimización iterativos. Es por esta razón que se comentan otros métodos de ajuste en principio motivados por ideas intuitivas, y que han demostrado tener mayor estabilidad numérica que los métodos basados en verosimilitud. Dichos métodos se enmarcan en el contexto de los estimadores minimizadores de distancia. En el resto de esta sección se supone una muestra dada de tamaño n de (I, C, L). Por simplicidad, se denotará a los elementos de dicha muestra por $(\mathbf{i}_k, \mathbf{c}_k, \mathbf{l}_k)$ con $k \in \{1, \ldots, n\}$. Se espera esta notación no cause confusión con respecto al soporte de (I, C, L), descrita en la Sección 3.2.

Para estimar el parámetro de la distribución de I se considera el estimador usual de una variable aleatoria $\mathrm{Ber}(p)$, $\hat{p} = n_1/n$ donde $n_1 = \sum_{k=1}^{n} \mathbf{1}_1(\mathbf{i_k})$. Para los parámetro restantes se proponen las alternativas de las siguientes secciones.

4.1.1. Estimación por máxima verosimilitud

Del desarrollo en la Sección 3.2 del Capítulo 3, la probabilidad conjunta de i, c, l, o su función de verosimilitud, es

$$\mathbb{P}\left(I = i, C = c, L = l\right) = \mathbb{P}\left(I = i\right)\mathbb{P}\left(C = c | I = i\right)\mathbb{P}\left(L = l | C = c\right). \tag{4.1}$$

Ahora, para conocer la relación funcional de (4.1), se escribe cada uno de los términos según las distribuciones especificadas en la Sección 3.4.1 del Capítulo 3. El primer término está dado por

$$\mathbb{P}(I = i) = p^{\mathbf{1}_1(i)}(1-p)^{\mathbf{1}_0(i)}. \tag{4.2}$$

Para los términos siguientes se supondrá que los catálogos están ordenados de mayor a menor con respecto al valor de su función de probabilidad correspondiente. Más aún, se

empleará la notación presentada en la Sección 3.2 del Capítulo 3: c_k^i denota el k-ésimo contig del individuo i. A su vez, l_m^c denota la m-ésima lectura proveniente del contig c. Además, se consideran las funciones

$$f_1(\alpha, k) = g(\alpha, k) - \log\left(\sum_{j=1}^{n_{\mathcal{C}(i)}} \exp\{g(\alpha, j)\}\right) \tag{4.3}$$

siendo $g(\alpha, m) = -\alpha_1 \log(\alpha_2 m) + \alpha_3 m / n_{\mathcal{C}(i)}$, y

$$f_2(\theta_c, k) = \theta_c^k / \left(\sum_{j=1}^{n_{\mathcal{L}(c)}} \theta_c^j\right). \tag{4.4}$$

Entonces, con la notación anterior,

$$\mathbb{P}\left(C = c | I = i\right) = \prod_{k=1}^{n_{\mathcal{C}(i)}} \left(\exp\{f_1(\alpha_i, k)\}\right)^{\mathbf{1}_{c_k^i}(c)}. \tag{4.5}$$

Finalmente, el último término en (4.1) está dado por

$$\mathbb{P}\left(L = l | C = c\right) = \prod_{m=1}^{n_{\mathcal{L}(c)}} \left(f_2(\theta_c, k)\right)^{\mathbf{1}_{l_m^c}(l)}. \tag{4.6}$$

De todo lo anterior, sustituyendo los términos de (4.2), (4.5), (4.6) en (4.1), se obtiene que la log verosimilitud está dada por

$$\begin{aligned}
&\mathrm{lv}(i, c, l; p, \alpha, \theta) = \\
&\mathbf{1}_1(i)\log(p) + \mathbf{1}_0(i)\log(1-p) + \sum_{k=1}^{n_{\mathcal{C}(i)}} \mathbf{1}_{c_{ik}}(c) f_1(\alpha_i, k) + \sum_{m=1}^{n_{\mathcal{L}(c)}} \mathbf{1}_{l_{cm}}(l)\log(f_2(\theta_c, m)).
\end{aligned}$$

Ahora, al derivar la función en (4.1.1), los términos de la función Score están dados por

$$\frac{\partial \mathrm{lv}}{\partial p}(i, c, l; p, \alpha, \theta) = \frac{\mathbf{1}_1(i)}{p} - \frac{\mathbf{1}_0(i)}{1-p}, \tag{4.7}$$

$$\frac{\partial \mathrm{lv}}{\partial \alpha}(i, c, l; p, \alpha, \theta) = \sum_{k=1}^{n_{\mathcal{C}(i)}} \mathbf{1}_{c_{ik}}(c) \frac{\partial f_1}{\partial \alpha}(\alpha_i, k), \tag{4.8}$$

$$\frac{\partial \mathrm{lv}}{\partial \theta}(i, c, l; p, \alpha, \theta) = \sum_{m=1}^{n_{\mathcal{L}(c)}} \mathbf{1}_{l_{cm}}(l) \frac{1}{f_2(\theta_c, m)} \frac{\partial f_2}{\partial \theta}(\theta_c, m). \tag{4.9}$$

Una primera observación es que para el caso de la distribución del individuo, el estimador vía máxima verosimilitud coincide con el que se ha propuesto en la sección anterior. Por otra

parte, los estimadores para α y θ pueden encontrarse de manera paralela. El reto principal al optimizar se debe a que el término correspondiente al contig más abundante o a la lectura más abundante será el que tenga mayor influencia en la función. Lo anterior provocará —como ha sido corroborado de manera empírica por el autor— que el ajuste para los elementos menos abundantes no sea adecuado. Por lo tanto es recomendable estimar los parámetros con una versión regularizada de la log verosimilitud. En desarrollos posteriores convendría estudiar e implementar con mayor profundidad los estimadores de máxima verosimilitud obtenidos para α y θ y cómo abordar el inconveniente mencionado.

A continuación se presenta otro método para estimar los parámetros del modelo que se basa en la minimización de cierta distancia entre la función de distribución propuesta y la empírica. Una fuente bibliográfica para el análisis de estimaciones semejantes a la propuesta descrita enseguida —pero tratada para casos específicos— puede encontrarse en William (1981). Se trata de una gran clase de estimadores conocida como estimadores de mínima distancia.

4.1.2. Estimación por mínima distancia

Se proponen métodos alternativos a los estimadores de máxima verosimilitud. Se comienza proponiendo un método para la estimación del parámetro de la distribución de $C|I = i$. Primero se calcula el vector $\mathbf{p}_{\text{est}} = (\hat{p}_{ic_k})_{k\in\{1,\dots,n_{\mathcal{C}(i)}\}} \in [0,1]^{n_{\mathcal{C}(i)}}$ siendo $n_{\mathcal{C}(i)}$ la cardinalidad de $\mathcal{C}(i)$ y $\hat{p}_{ic_k} = n_{ic_k}/n_i$ con n_i definido de manera análoga a n_1 y $n_{ic_k} = \sum_{s=1}^{n} \mathbf{1}_{(i,c_k)}(\mathbf{i}_s, \mathbf{c}_s)$. Posteriormente, se ordenan de mayor a menor las entradas de $\mathbf{p}_{\text{est}}$ y, por abuso de notación, se llama $\mathbf{p}_{\text{est}}$ al vector resultante al ordenar de esta manera las entradas. A continuación se considera la función f_1 definida en (4.3). Finalmente se obtiene de manera numérica el argumento α que minimiza

$$\frac{1}{n_{\mathcal{C}(i)}} \sum_{j=1}^{n_{\mathcal{C}(i)}} \left[f_1(\alpha, j) - \log\{(\mathbf{p}_{\text{est}})_j\} \right]^2 . \tag{4.10}$$

Con el objetivo de analizar empíricamente la calidad de este ajuste, a continuación se presentan los resultados obtenidos al aplicar el método propuesto. Se ha empleado la función

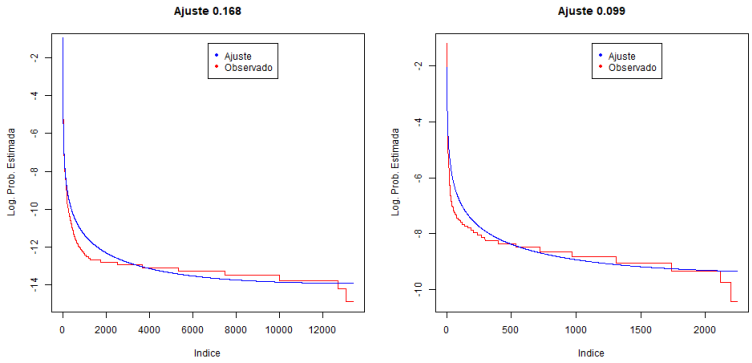

Figura 4.1: Ajuste del modelo propuesto para la distribución de $C|I$ en escala log. El ajuste para el parásito se observa en la imagen de la derecha.

`solnl` de la paquetería `NlcOptim` (Chen & Yin, 2017) del software estadístico `R` (Ihaka & Gentleman, 1996). Primero se muestra en la Tabla 4.1 el valor numérico de los parámetros por individuo.

Tabla 4.1: Estimaciones de los parámetros de la distribución de los contigs dado el individuo para el huésped y para el parásito. Se observa una diferencia notable entre los parámetros del huésped y los del individuo. El valor de α_2 se redondeó a 1 pues al realizar el ajuste este no cambia.

Individuo/Parámetro	α_1	α_2	α_3
Huésped	1.527	1	1.527
Parásito	1.047	1	0.781

Al observar la Tabla 4.1, se concluye que α_2 es un parámetro que en desarrollos posteriores podría ser omitido. En este caso el valor obtenido ha sido muy cercano a 1 y por lo tanto se ha redondeado a tal valor. Por otra parte, los parámetros restantes reflejan la diferencia observada entre las curvas correspondientes para cada individuo. Con fines ilustrativos, se presentan en la Figura 4.1 algunos de los ajustes obtenidos.

4.1.3. Estimación empleando norma $||\cdot||_{L_1}$

Finalmente se propone otro método de ajuste para las distribuciones condicionales $L|C = c$. Se hace de manera análoga al caso anterior pero, con la intención de mostrar los puntos donde difiere, es desarrollado nuevamente. Primero se calcula el vector $\mathbf{p}_{\text{est}} = (\hat{p}_{cl_k})_{k\in\{1,\dots,n_{\mathcal{L}(c)}\}} \in [0,1]^{n_{\mathcal{L}(c)}}$ siendo $n_{\mathcal{L}(c)}$ la cardinalidad de $\mathcal{L}(c)$ y $\hat{p}_{cl_k} = n_{cl_k}/n_c$ con $n_c = \sum_{j=1}^{n} \mathbf{1}_c(\mathbf{c}_j)$ y $n_{cl_k} = \sum_{s=1}^{n} \mathbf{1}_{(c,l_k)}(\mathbf{c}_s, \mathbf{l}_s)$. Posteriormente se ordenan de mayor a menor las entradas de $\mathbf{p}_{\text{est}}$ y, por abuso de notación, se llama $\mathbf{p}_{\text{est}}$ al vector resultante al ordenar de esta manera las entradas. A continuación se considera la función f_2 definida en (4.4). Finalmente se estima de manera numérica el $\theta_c \in [0,1]$ que minimiza

$$\frac{1}{n_{\mathcal{L}(c)}} \sum_{j=1}^{n_{\mathcal{L}(c)}} \left| f_2(\theta_c, j) - (\mathbf{p}_{\text{est}})_j \right|. \tag{4.11}$$

Como ejemplo de los resultados obtenidos al realizar estos ajustes se presentan los que se encuentran en la Figura 4.2. En este caso se ha empleado la función `optimize` de `R`. Con

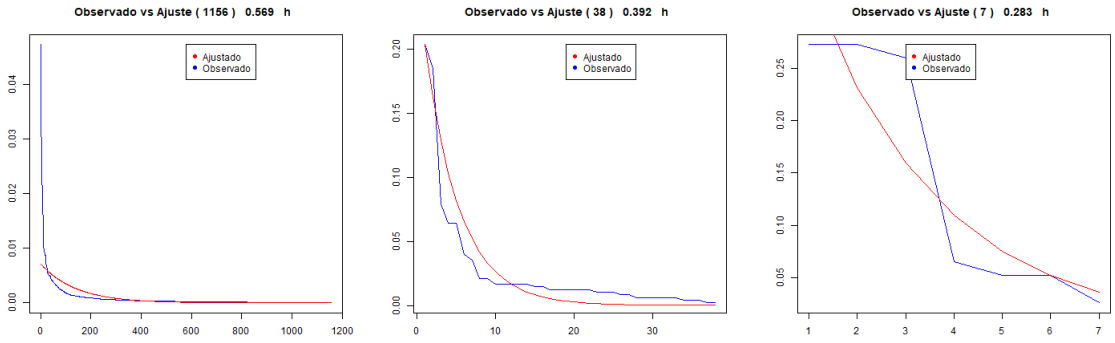

Figura 4.2: Muestra en distintos contigs del ajuste con la distribución cuyo kernel es θ_c^k. En azul se grafica la proporción observada y en rojo la curva ajustada.

el objetivo de observar cómo son los parámetros estimados, se ha realizado una estimación de su densidad —simplemente como recurso visual y sin suponer que estos tienen cierta distribución— y se ha separado por individuo. Los resultados se muestran en la Figura 4.3. Bajo el modelo propuesto se esperaría que estas curvas no difirieran entre sí. Sin embargo, se

observa que sí existe diferencia entre ellas. Una posible explicación es que en el parásito la distribución de los parámetros estimados cambiarán con relación al tiempo que se deje correr el experimento debido a la degradación de los fragmentos de ARN [1]. De la Figura 4.2 se ob-

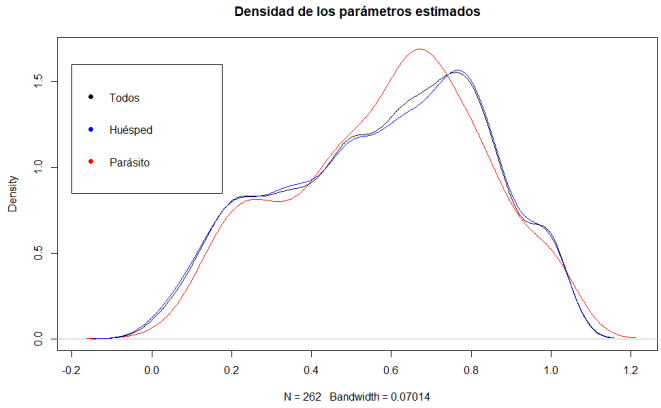

Figura 4.3: Distribución de los parámetros estimados. A pesar de que se esperaría ver que la distribución de los parámetros no depende del individuo esto no es así. Lo anterior sugiere que para desarrollos posteriores podría considerarse dependencia del individuo en la distribución de la lectura.

serva que en algunos casos el ajuste del modelo es insuficiente para describir ciertas sutilezas. Se puede tomar una distribución más compleja tal como la propuesta para la distribuciones condicionales $C|I = i$ obteniendo así un mejor ajuste al tratarse de un modelo flexible. Esto es, se sugiere emplear la función f_1 en lugar de la función f_2. Sin embargo, ahora el análisis de los parámetros y el proceso de optimización es un poco más elaborado. Si es de interés estimar la probabilidad las lecturas más abundantes es recomendable emplear el modelo más complejo.

4.2. Simulación de datos

Finalmente, con el objetivo de estudiar empíricamente la adecuación del modelo a la situación de estudio, se han simulado datos recurriendo a la estructura que el modelo espe-

[1] Lo anterior es sólo una conjetura.

cifica. La meta es generar datos artificiales y observar si estos presentan un comportamiento cualitativo semejante al observado en los datos proporcionados.

El esquema de simulación es el siguiente:

1. Simular I acorde a una variable Bernoulli de parámetro p.
2. Simular un C acorde al modelo de abundancia dependiente del valor de I.
3. Simular una lectura L acorde al modelo de abundancia del contig C.

Como se observó al explorar los datos, es posible que existan entre los contigs cierta variabilidad respecto a al adecuación del modelo de abundancia específico. Por lo tanto, en el paso 3 es posible realizar una modificación que tome en cuenta esta variabilidad. Más aún, con fines de simulación es recomendable simular en el paso 3 acorde al mismo modelo que para el de los contigs dado el individuo. Lo anterior se debe a que, por cuestiones numéricas, para el modelo con kernel θ^k, algunas probabilidades estimadas pueden ser 0 en la práctica, mientras que en los datos no es así. Lo dicho se ilustra al comparar en la Figura 4.4 el ajuste con cada uno de los modelos para el contig 66.

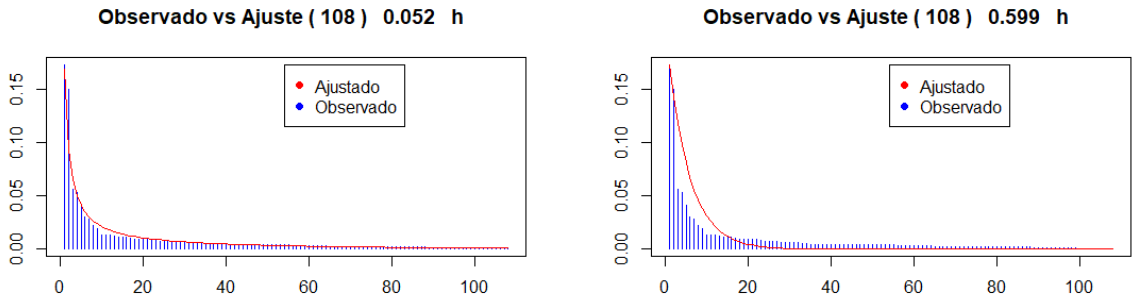

Figura 4.4: Comparación del comportamiento de dos propuestas de distribuciones distintas que se pueden considerar para el paso 3. Se observa que el modelo de la izquierda es adecuado para reflejar las sutilezas de la distribución observada.

Como primer ejercicio de simulación se ha especificado el parámetro de la distribución para el individuo igual a 0.015. Para las distribuciones basadas en modelos de abundancia

se han tomado los parámetros estimados anteriormente. Se obtienen 10^6 observaciones de datos simulados. En este escenario de simulación no existe ambigüedad en la asignación de las lecturas debida a la compatibilidad con dos contigs. Sin embargo, aún está presente la ambigüedad derivada por los contigs que están asociados a ambos individuos. En la Figura 4.5 se muestra la comparación —para diversos contigs— entre las frecuencias simuladas y las frecuencias observadas. Como era de esperarse, la comparación entre los datos simulados

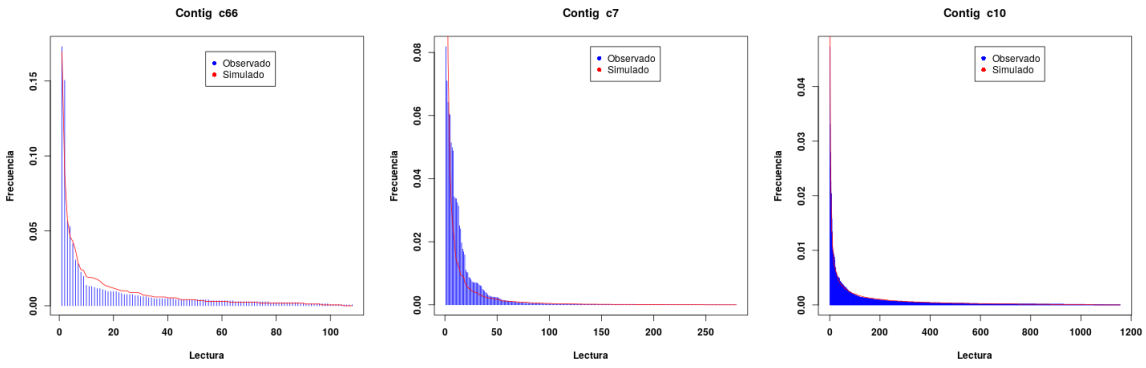

Figura 4.5: Comparación entre datos observados y simulados. En azul se muestra la frecuencia relativa observada de la lectura en el contig en cuestión. En rojo se muestra la frecuencia relativa de la lectura obtenida al simular los datos.

y observados es parecida a la comparación entre el ajuste teórico y los datos observados. Ahora bien, un segundo ejercicio de simulación consiste en comparar la procedencia real de las lecturas simuladas y la que sería observada.

Contar con este esquema de simulación permitirá, en desarrollos posteriores, producir un conjunto de datos sintéticos en los cuales se conocen sin lugar a dudas los parámetros y la clasificación correcta. Entonces, con este tipo de datos se pueden realizar tareas de comparación con otras técnicas o métodos de clasificación que se propongan o empleen actualmente.

CAPÍTULO 5

Aportes del modelo: solución a problema de clasificación y discusión.

En este último capítulo se presenta una propuesta de solución —elaborada con base en los elementos especificados en los capítulos anteriores— al problema de asignación planteado inicialmente. Posteriormente se comentan algunos alcances del modelo propuesto. Finalmente se dan conclusiones generales sobre la tesis desarrollada.

5.1. Propuesta de solución vía variante estocástica del algoritmo EM

A continuación se especifica la manera en que puede ser aplicado el algoritmo EM en este contexto. Como se verá al plantearlo, una vez que se ha postulado un modelo y una técnica de estimación para sus parámetros, la aplicación es casi directa. Los pasos a seguir son los siguientes:

1. Estimar los parámetros del modelo —para obtener un valor inicial— con una asignación dada. Una manera de realizar lo anterior es estimarlos de manera análoga a como se hizo en la Sección 4.1 del Capítulo 4.

2. Para cada lectura ambigua l

   a) calcular los estimados de $(\mathbb{P}(I = i, C = c, L = l))_{(i,c)\in \mathrm{Comp}(l)}$ donde $\mathrm{Comp}(l)$ es el conjunto de pares de individuos y contigs a los cuales puede ser asignada la lectura l,

   b) asignar las repeticiones de l según la realización de una variable aleatoria multinomial de parámetro n_l —el número de repeticiones de l— clases los elementos en $\mathrm{Comp}(l)$ y probabilidades las estimadas anteriormente. Es de notar que este aspecto difiere el algoritmo EM con el empleado comúnmente (ver Li et al. (2009)). La razón es que en la versión clásica los valores estimados serían sustituidos en la expresión de cierta esperanza condicional.

3. Con esta nueva asignación volver a estimar los parámetros del modelo.

4. Repetir desde el paso 2 hasta satisfacer un criterio de paro tal como número de iteraciones, tolerancia en el cambio de los estimados de los parámetros, etc.

5.1.1. Aplicación de propuesta en datos en Bermudez-Barrientos et al. (2019)

Con el fin de ilustrar la aplicación de la propuesta de solución en los datos de Bermudez-Barrientos et al. (2019), ahora se presenta un análisis más detallado de las lecturas ambiguas y la clasificación de las repeticiones de una lectura en particular. En el resto de esta sección se empleará la notación de los datos proporcionados y derivados del experimento de Bermudez-Barrientos et al. (2019) mencionado en el Capítulo 2.

Primero se obtiene cuántas lecturas únicas ambiguas hay y de qué tipo. Los tipos propuestos son similares a los considerados en Li et al. (2009). Se dirá que una lectura es ambigua

del tipo 0 si está asociada a un solo contig y a un solo individuo. Si una lectura ambigua está asociada a más de un contig pero sólo a un individuo, será ambigua del tipo 1. Finalmente, si una lectura está asociada a más de un individuo, se dirá que es del tipo 2. La tipología anterior está basada en los términos del modelo probabilístico que sería necesario tomar en cuenta para clasificar las repeticiones de las lecturas. En la Tabla 5.1 se muestra cuántas lecturas hay por tipo, así como el número máximo de repeticiones, y el número de lecturas con una sola repetición.

Tabla 5.1: Lecturas únicas y algunas de sus características por tipo de ambigüedad. El tipo de ambigüedad más presente es el del tipo 1. Además es en el tipo para el cuál existe una lectura con un número de repeticiones considerable (8511). Por tipo 0 se está refiriendo a que la lectura no presenta ambigüedad.

Tipo	0	1	2
Número de lecturas	28400	177	94
Número máximo de repeticiones	696756	8511	215
Lecturas únicas con una sola repetición	1476	15	9

Para las lecturas que sólo cuentan con una repetición hay poco qué hacer respecto a su clasificación. Debido a que al asignarla a un contig automáticamente se eliminará del catálogo de cualquier otro al que esté asociada. Es decir, después de la primera iteración del algoritmo EM, éstas dejarán de ser consideradas como ambiguas. Ahora bien, como se puede observar en la Tabla 5.1, son relativamente pocas las lecturas ambiguas que presentan tal situación.

Ahora se muestra el ejemplo particular de la asignación de una lectura ambigua del tipo 1 en la primera iteración del algoritmo EM. Como primer paso para aplicar el algoritmo EM es necesario contar con una estimación inicial de los parámetros del modelo. En este ejemplo se han tomado como parámetros iniciales los estimados en la Sección 4.1.

La lectura ambigua a analizar es la lectura “s1021”. Dicha lectura cuenta con 138 repeticiones y es la lectura que “rompía” el patrón, mencionada en la Sección 3.3 del Capítulo 3. Ésta se encuentra asociada con los contigs c1 y c4. Al ser s1021 ambigua del tipo 1, los

Tabla 5.2: Probabilidad condicional estimada para los contigs dado el individuo. Basados en estas probabilidades únicamente, la mayoría de las repeticiones serían asignadas al contig c1.

c	c1	c4
$\mathbb{P}(C = c \mid I = 0)$	0.4	0.05

Tabla 5.3: Probabilidad condicional estimada para la lectura dados los contigs. Si sólo tomásemos en cuenta esta probabilidad condicional, ahora la mayoría de las repeticiones serían asignadas al contig c4.

c	c1	c4
$\mathbb{P}(L = s1021 \mid C = c)$	1.7e-04	1e-03

contigs c1 y c4 están asociados con el mismo individuo, a saber, están asociados al huésped. En la Tabla 5.2 se muestra la probabilidad condicional estimada para cada uno de los contigs. Siguiendo el enfoque heurístico se tendría que, de las 138 lecturas, serían asignadas al contig c1, aproximadamente, 72[1]. Por otra parte, se verá que con el enfoque de EM no es así. En la Tabla 5.3 se encuentra el valor de la distribución condicional de la lectura con respecto a cada contig. Lo reflejado en la Tabla 5.3 es que, intuitivamente, la lectura s1021 no es tan importante para c1 como lo es para c4. Formalmente, esto se ve reflejado en la probabilidad condicional —en la cual se basaría la asignación de las repeticiones de la lectura– que en este caso mostramos en la Tabla 5.4.

Así, asignando según la realización de una variable aleatoria con distribución multino-

[1] Pues el contig c1 cuenta con 28 lecturas únicas no ambiguas y el contig c4 con 26.

Tabla 5.4: Probabilidad condicional de los contigs dada la lectura. En el esquema de clasificación empleando EM, la asignación de las repeticiones se basa en la probabilidad condicional mostrada.

c	c1	c4
$\mathbb{P}(C = c \mid L = s1021)$	0.58	0.42

mial (paso 2.b del algoritmo presentado en la Sección 5.1), el valor esperado de repeticiones asignadas al contig c1 sería aproximadamente 80, mientras que el valor esperado de repeticiones asignadas al contig c4 sería 58. Con esta asignación no se descarta la posibilidad de asignar 138 repeticiones al contig c1 y 0 al contig c4. Si se da ese caso la lectura dejaría de ser ambigua y sería necesario actualizar los catálogos de lecturas y contigs. Por otra parte, es de esperar que al iterar el algoritmo, el número de repeticiones asignadas al contig c4 aumente pues la probabilidad de la lectura dado el contig c1 disminuirá al tener menos repeticiones asignadas a éste.

Notemos que algoritmicamente es relativamente sencillo de plantear el uso de la variante de EM. Sin embargo, la implementación y el manejo eficiente de la base de datos son situaciones que potencialmente complican su aplicación. Por ejemplo, un reto importante es la actualización del catálogo en caso de que las lecturas dejen de ser ambiguas. i.e., que todas sus repeticiones estén asignadas a un solo contig e individuo.

5.1.2. Nota sobre implementación de algoritmo EM en este contexto

Como ha sido mencionado, el principal obstáculo radica en el manejo eficiente de los datos. Se propone entonces una estructura simple que da solución a este problema a costa del uso de memoria. Esta estructura de datos está inspirada en la proporcionada por el Dr. Cei Abreu Goodger.

Se considera un vector M_h para el huésped y otro M_p para el parásito. Estos vectores tendrán tantas entradas como pares del tipo (l, c) donde la lectura l puede ser asociada al contig c. i.e., $M_h \in (\mathbb{N})^{n_{\mathcal{A}(0)}}$ y $M_p \in (\mathbb{N})^{n_{\mathcal{A}(1)}}$ donde $n_{\mathcal{A}(i)}$ es la cardinalidad del conjunto $\{(l, c) : l \text{ es una lectura asociada al contig } c \text{ está asociado al individuo } i\}$.

Ahora bien, inicialmente, las entradas de estos vectores se rellenan con los datos con que se cuenta originalmente. Es decir, las repeticiones de lecturas únicas ambiguas se consideran provenientes de todos los contigs e individuos a las cuáles están asociadas.

Posteriormente, se estiman los parámetros del modelo como se ha descrito en el Capítulo 4. En este paso es necesario tomar en cuenta que, para los contigs que pueden provenir tanto

del parásito como del huésped, las repeticiones de las lecturas serán la suma de las entradas correspondientes en los vectores, i.e., si c es un contig que puede ser asociado tanto al huésped como al parásito, el número de repeticiones de lecturas asociado a c será

$$\sum_{l \in \mathcal{L}(c)} (M_h)_{(l,c)} + \sum_{l \in \mathcal{L}(c)} (M_p)_{(l,c)},$$

recordando que $\mathcal{L}(c)$ denota las lecturas que pueden provenir del contig c. La misma observación vale para las lecturas asociadas a múltiples individuos. Las repeticiones de l asociadas al contig c serán $(M_h)_{(l,c)} + (M_p)_{(l,c)}$.

Posteriormente, para cada lectura única ambigua l —que, como lo muestra la Tabla 5.1, son 271— se realiza la asignación de sus repeticiones de manera análoga al ejemplo mostrado en la Sección 5.1 y se actualizan las entradas correspondientes en los vectores M_h y M_p. El procedimiento anterior se repite hasta que se satisfaga un criterio de convergencia.

Finalmente, cabe resaltar que la implementación anterior aún está en abstracto pues para el manejo computacional de los vectores M_h y M_p —es decir, acceso en memoria a sus entradas y su actualización— existen varias alternativas y estas difieren en eficiencia. También es de observar que, al ser 271 las lecturas únicas ambiguas, es posible manejar estructuras distintas para las lecturas únicas ambiguas y para las no ambiguas. En la práctica, sólo se considera el último tipo de lecturas para realizar la estimación de los parámetros.

5.1.3. Resultados obtenidos para datos en Bermudez-Barrientos et al. (2019)

Los resultados presentados en este apartado son preliminares y aún es necesario llevar a cabo un análisis con mayor profundidad. Tal análisis contempla la determinación de las condiciones para que las estimaciones de los parámetros del modelo converjan, bondad de ajuste del modelo y complejidad computacional. Se mencionan los retos afrontados, los retos por afrontar y algunas vías para realizarlo.

La implementación se realizó en el software `R` y se ha empleado las paqueterías `NlcOptim` y `plyr` (Anderson, 2012). La primera paquetería está diseñada para resolver numéricamen-

te problemas de optimización no lineales con restricciones. La segunda paquetería permite realizar operaciones semejantes a las que se realizan utilizando el lenguaje SQL.

Para obtener los resultados mostrados se ha fijado el número de iteraciones igual a 100 y se han tomado como parámetros iniciales los obtenidos en el Capítulo 4. El valor estimado para el parámetro de la distribución de I es 0.01121. En la Figura 5.1 se muestran los cambios en el parámetro con respecto al número de iteración. Por su parte, para la distribución de los contigs, los valores estimados son los presentados en la Tabla 5.5.

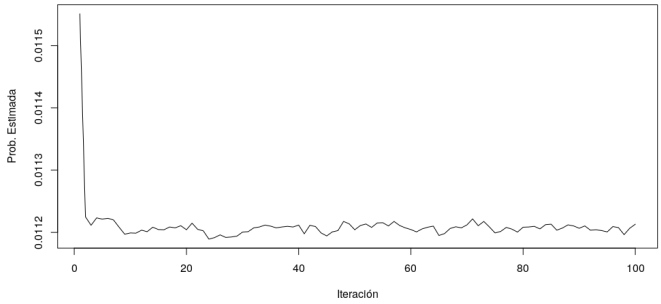

Figura 5.1: Valor del parámetro estimado para la distribución del individuo en distintas iteraciones. Se observa cierta estabilidad en las últimas iteraciones.

Tabla 5.5: Estimaciones de los parámetros de la distribución de $C|I$ por individuo al aplicar EM. Se observa una diferencia notable entre los parámetros del huésped y los del individuo. Comparar con Tabla 4.1

Individuo/Parámetro	α_1	α_2	α_3
Huésped	1.496	1	1.496
Parásito	0.988	1	0.929

Ahora, para los contig c tales que una lectura ambigua está asociada, se presentan en la Figura 5.2 los parámetros obtenidos en distintas iteraciones. Se observa que para algunos contigs, las estimaciones parecen converger. En contraste, para gran parte de los contigs, los parámetros estimados varían considerablemente de iteración a iteración. Lo anterior plantea la pregunta importante sobre técnicas para obtener estimadores puntuales.

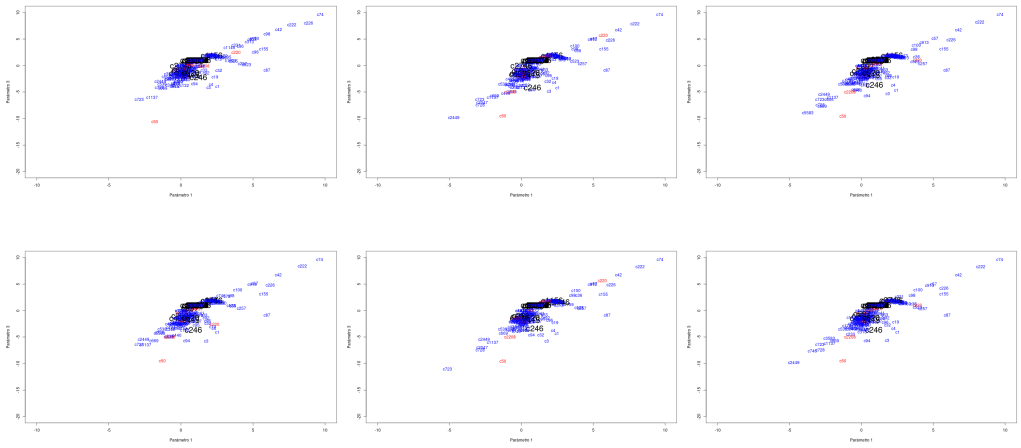

Figura 5.2: Ejemplos del valor estimado de los parámetros de la distribución de $L|C$, para distintos contigs, en distintas iteraciones. La convergencia de los parámetros estimados sólo es clara para unos cuantos contigs. En azul se muestran los contigs asociados al huésped, en rojo los asociados al parásito, y en negro los ambiguos.

5.2. Comentarios finales

- Con fines de evaluar la calidad en el sentido estadístico de los estimadores propuestos, es de interés estudiar sus propiedades asintóticas. Al respecto, un artículo que provee de referencias para problemas similares es William (1981). Más aún, la regularidad misma del modelo es un aspecto a estudiar con mayor profundidad.

- Al haberse propuesto un modelo con muchos menos parámetros que los considerados en modelos estándard (Li et al. (2009), Salzman et al. (2011)), es natural preguntarse si esto posibilita el diseño de pruebas de hipótesis más simples que las actuales. Es bien sabido que al analizar expresión genética, un problema que surge es el de pruebas de hipótesis múltiples.

- El autor considera, y esta es una de las razones de haber optado por la versión presentada de EM, que computacionalmente existen ventajas al aplicar esta variante de EM y no la clásica considerada en Li et al. (2009). La principal ventaja radica en que las esti-

maciones de los parámetros se pueden realizar de manera paralela. Por otra parte, una pregunta aún por responder es qué técnica es conveniente para proveer de estimadores puntuales de los parámetros.

- A pesar de que éste es un modelo *ad hoc* para la situación específica del experimento en Bermudez-Barrientos et al. (2019), conviene averiguar si el modelo es lo suficientemente flexible para adecuarse a otras situaciones similares.

- Siguiendo con el ejemplo presentado para ilustrar el funcionamiento del algoritmo EM en este caso, es posible desarrollar una herramienta de software donde se seleccionen lecturas ambiguas y paso a paso se muestren los estimados de las probabilidades condicionales en cuestión junto al gráfico de éstas, proveyendo una herramienta didáctica para usuarios potenciales del modelo propuesto.

- Los contigs no son el único catálogo que se puede tener en cuenta. Existe la posibilidad de considerar otros catálogos y es de interés para trabajo futuro averiguar si el modelo es lo suficientemente flexible para realizarlo. Haciendo referencia a los resultados presentados en los ajustes, conviene estudiar la contribución del uso de modelos de abundancia de especies para el estudio de expresión genética.

- Este es un problema que puede enmarcarse en el contexto de aprendizaje supervisado pero con la peculiaridad de que existe información adicional que podría ser independiente al modelo de probabilidad: la compatibilidad de las lecturas con ciertos contigs, y de contigs con individuos. No se trata del contexto común de aprendizaje supervisado en el cual el conjunto de entrenamiento por sí mismo permite estimar los parámetros del modelo. En efecto, por definición de ambigüedad, algunos puntos son imposibles de clasificar correctamente para fines de entrenar al modelo de aprendizaje. Más aún, al considerar un catálogo fijo, la estimación requiere de la inclusión de los elementos cuya clasificación se desconoce. En este sentido, puede concebirse como una mezcla de aprendizaje no supervisado y aprendizaje supervisado. Lo anterior sugiere un posible tema de reflexión respecto a identificar si se trata de un planteamiento teórico

innovador.

- En el desarrollo del modelo propuesto no se han empleado de manera literal los modelos de abundancia pues estrictamente estos modelan los conteos obtenidos. Lo que se ha modelado de esta manera son los valores de las funciones de probabilidad condicionales. Sin embargo, para desarrollos y análisis posteriores —donde el interés sea sólo los conteos en los contigs y no así un modelo que permita realizar clasificación— los modelos de abundancia en su más pura versión pueden ser empleados.

- Finalmente, vale la pena hacerse la pregunta por la diferencia cualitativa entre estimar los parámetros con las lecturas ambiguas y con la clasificación realizada. Es decir, si las conclusiones obtenidas variarían considerablemente. Con el modelo de probabilidad y el esquema de simulación propuesto se puede comenzar con dicha investigación. Por su parte, para abordar la pregunta sobre la diferencia cuantitativa, la propuesta del modelo ya es un avance de importancia en sí misma.

APÉNDICE A

Herramientas de probabilidad y estadística.

Este apéndice presenta conceptos propios de probabilidad y estadística mencionados en el texto.

A.1. Modelos gráficos dirigidos

El material presentado en esta sección se basa en Jordanet al. (2004) y nos limitaremos al caso de los modelos gráficos probabilísticos dirigidos. En varios contextos de aplicación es de interés modelar el comportamiento de más de una variable en cierto sistema. Ahora bien, es común que la distribución de ciertas variables se conozca de mejor manera o sea más fácil de entender condicionada a que se ha observado un valor específico de otra variable. El objetivo de los modelos gráficos —o de los modelos gráficos probabílisticos— es proveer de herramientas formales para modelar de esta manera: especificando la distribución conjunta de las variables tratadas al especificar sus distribuciones condicionales y relaciones de dependencia.

Presentamos el desarrollo para el caso de variables aleatorias discretas y de modelos gráficos dirigidos. Dada una sucesión de espacios finitos discretos $(\mathcal{X}_v)_{v\in\mathcal{V}}$ con $\mathcal{V}$ un conjunto fini-

to de elementos y una gráfica $\mathcal{G}(\mathcal{V}, E)$ dirigida acíclica, se especifica la distribución conjunta de $(x_v)_{v\in\mathcal{V}}$ mediante

$$f\left((x_v)_{v\in\mathcal{V}}\right) = \prod_{v\in\mathcal{V}} k_v(x_v, x_{\pi_v}) \tag{A.1}$$

donde π_v denota el conjunto de padres del nodo v, $\{w : (w, v) \in E\}$, en la gráfica $\mathcal{G}$ y k_v es una función que satisface las siguientes propiedades:

1. $\sum_{x\in\mathcal{X}_v} k_v(x, x_{\pi_v}) = 1$.

2. $k_v(x, x_{\pi_v}) \geq 0$ para todo $x \in \mathcal{X}_v$.

Con dichas propiedades, f definida como en (A.1) es una función de distribución en $\prod_{v\in\mathcal{V}} \mathcal{X}_v$ y que además, si $(X_v)_{v\in\mathcal{V}}$ es una variable aleatoria con distribución f,

$$\mathbb{P}(X_v = x_v | X_{\pi_v} = x_{\pi_v}) = k_v(x_v, x_{\pi_v}). \tag{A.2}$$

Las funciones k_v que se suelen tomar son aquellas funciones que se desean especificar como las probabilidades condicionales de interés. De la formulación anterior se pueden derivar otras propiedades tales como las de variables condicionalmente independientes.

A.2. Modelos de abundancia de especies

La exposición presentada en esta sección se basa en McGill et al. (2007). Los modelos de abundancia de especies son una gama de modelos para la distribución de abundancia de especies en un ecosistema determinado. Estos se explican de mejor manera partiendo de los datos observados.

Supongamos un ecosistema que consta de un conjunto de individuos y un conjunto de clases disjuntas $C_1, \ldots, C_k$ a las cuales puede pertenecer cada individuo. Si $\mathbf{N} = (N_1, \ldots, N_k)$ es el vector con número de individuos observados de cada clase, al ordenar las entradas de $\mathbf{N}$ de mayor a menor suele surgir un patrón claro: son pocas las clases para las cuales el número de individuos es bastante grande y las demás clases cuentan con un número de individuos

bastante menor y casi constante entre éstas. Para emular este comportamiento, en la literatura se han propuesto distribuciones de conteo que están relacionados de alguna manera.

Un ejemplo de tales modelos es el lognormal-Poisson. Este postula que cada clase contiene un número de individuos distribuidos como una variable aleatoria Poisson con intensidad la realización de una misma variable —para todas las clases en cuestión— distribuida como una lognormal. Dependiendo del modelo en cuestión se pueden plantear las ecuaciones de verosimilitud para estimar los parámetros de interés.

A.3. Algoritmo EM

El algoritmo EM es un algoritmo ampliamente conocido en el área de estadística, especialmente para la obtención de los estimadores de máxima verosimilitud en el caso de datos faltantes. Es con este objetivo que se presenta en Dempster et al. (1977). Su formulación clásica es la siguiente.

Supongamos que se tienen dadas variables aleatorias X, Y con función de verosimilitud $f(x, y; \theta)$ con θ un parámetro a estimar. Dada una muestra $x_1, \ldots, x_n$ nuestro interés es estimar θ. Se procede entonces de la siguiente manera:

1. Calcular la esperanza condicional

$$Q(\theta, \theta_0) = \mathbb{E}\left(\sum_{i=1}^{n} \log f(Y_i, X_i; \theta) | X_i = x_i\right) \tag{A.3}$$

   donde se supone conocido y dado θ_0.

2. Maximizar $Q(\theta, \theta_0)$ con respecto a θ y hacer θ_0 igual al argumento obtenido.

3. Repetir los pasos anteriores hasta que se satisfaga un criterio de convergencia.

El nombre del algoritmo proviene de las siglas de los primeros dos pasos: Expectation y Maximization. Para el estudio de las propiedades de los estimadores obtenidos de esta manera se recomienda la lectura de Dempster et al. (1977) y literatura posterior.

Ahora bien, en el contexto de clasificación también es empleado el algoritmo EM al considerar la clase del dato como una variable no observada. En (Celeux & Govaert, 1992) se presentan dos variantes de este algoritmo. La primera variante presentada es la que se emplea en nuestro contexto. A continuación se explica ésta de manera breve. Al igual que en el algoritmo EM suponemos que se tienen dos variables aleatorias X, Y siendo Y discreta, finita y no observada. Se procede de la siguiente manera dada una muestra $(x_i)_{i=1}^n$.

1. Para cada valor k en el soporte de Y, y para cada x_i, calcular la probabilidad $\mathbb{P}(Y = k|X = x_i)$.

2. Para cada x_i, asignar la observación x_i a la clase k con probabilidad $\mathbb{P}(Y = k|X = x_i)$.

3. Obtener los estimadores para los parámetros del modelo[1].

4. Repetir los pasos anteriores hasta que se satisfaga un criterio de convergencia.

Esta última variante es la que se emplea en el Capítulo 5. La principal diferencia radica en los estimadores que se obtienen. En el caso presentado los estimadores se obtienen por medio de la minimización de una función distinta al negativo de la log verosimilitud, mientras que en la versión clásica los estimadores que se utilizan son los dados por máxima verosimilitud.

[1] En Celeux & Govaert (1992) se emplean los estimadores de máxima verosimilitud.

Referencias

Anderson, S. (2012). A quick introduction to plyr.

Bermudez-Barrientos, J. R., Ramirez-Sanchez, O., Chow, F. W., Buck, A. H., & Abreu-Goodger, C. (2019). Disentangling srna-seq data to study rna communication between species.

Celeux, G. & Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, *14*(3), 315–332.

Chen, X. & Yin, X. (2017). Nlcoptim: Solve nonlinear optimization with nonlinear constraints.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299–314.

Jiang, H. & Wong, W. H. (2009). Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, *25*(8), 1026–1032.

Jordan, M. I. et al. (2004). Graphical models. *Statistical Science*, *19*(1), 140–155.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2009). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, *10*(10), 995–1015.

Salzman, J., Jiang, H., & Wong, W. H. (2011). Statistical modeling of rna-seq data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *26*(1).

William, C. P. (1981). Minimum distance estimation: a bibliography. *Communications in Statistics-Theory and Methods*, *10*(12), 1205–1224.