USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

MÉTODOS ESTADÍSTICOS EN BIOINFORMÁTICA

Celia Domínguez Robles

Xullo, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

MÉTODOS ESTADÍSTICOS EN BIOINFORMÁTICA

Celia Domínguez Robles

Xullo, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Métodos estadísticos en bioinformática
Breve descrición do contido
O obxectivo deste traballo é coñecer os conceptos e métodos estadísticos que se empregan no campo da bioinformática, en particular, no estudio das secuencias de ADN.
Recomendacións
Outras observacións

Índice

Resumo

Neste traballo abordarase o estudo das secuencias de ADN empregando os métodos estadísticos como ferramenta fundamental. Estudaranse diversas cuestión relacionadas co proceso de ensamblado que forma parte da secuenciación dun xenoma. Previamente levarase a cabo o estudo dos conceptos de probabilidades e variables aleatorias, así como unha introdución dos procesos estocásticos, en concreto, dos procesos de Poisson, que son necesarios para modelizar o proceso de secuenciación. Finalmente, presentarase o estudo dun caso práctico relacionado co xenoma bacteriano, accedendo a bases de datos xenéticas e empregando un software especializado no ensamblado de secuencias.

Abstract

The current paper focuses on the study of DNA sequences, employing statistical methods as key tools. Diverse issues related to the assembly process associated with genome sequencing will be studied. Previously, concepts of probability and random variables will be reviewed, as well as an introduction to stochastic processes, specifically, Poisson processes, which are necessary for modelling the sequencing process. Lastly, a practical case study, related to the bacterial genome will be presented, by accessing genetic databases and using specialized software for sequence assembly.

Introdución

A ciencia avanza rapidamente e con cada descubrimento multiplícase a información. A bioinformática xorde da necesidade de interpretar a gran cantidade de datos biolóxicos que van medrando cos avances científicos. Como o seu nome indica, combina informática e bioloxía, empregando técnicas computacionais para almacenar, clasificar e analizar datos biolóxicos, especialmente, datos xenéticos. Para levar a cabo esta labor é imprescindible a estadística, sen a cal non sería posible avaliar os resultados dos experimentos nin extraer conclusións.

A estadística é a rama das matemáticas que estuda a variabilidade. A través dos métodos estadísticos os matemáticos contribuímos ao tratamento e interpretación dos datos obtidos por outros científicos. A maioría dos fenómenos biolóxicos non son deterministas, neles está presente a incerteza e precisan ser modelizados mediante probabilidades, co fin de medir dalgunha forma o que pode ocorrer. En particular, neste traballo abórdase o ensamblado de secuencias de ADN e como a estadística permite coñecer toda a información relacionada con este problema.

Nos tres primeiros capítulos desenvólvense os contidos preliminares co obxectivo de sentar unha base teórica na que se apoia o posterior estudo de secuencias de ADN. No primeiro capítulo introdúcense os conceptos esenciais sobre o ADN e a organización do material xenético, proporcionando o marco biolóxico no que se vai traballar. No segundo capítulo trátanse as probabilidades e variables aleatorias, necesarias para estudar os fenómenos aleatorios. No terceiro capítulo preséntanse os procesos estocásticos para modelizar a evolución dos procesos aleatorios que ocorren ao longo do tempo ou espazo, facendo fincapé nos procesos de Poisson.

No cuarto capítulo analízase o ensamblado das secuencias de ADN recalcando cal é o papel da estadística neste proceso. Finalmente, no capítulo cinco preséntase un caso práctico que ilustra o ensamblado de ADN, o que implica a obtención de datos xenéticos e o uso de ferramentas computacionais.

Tratarase de plasmar como a estadística é fundamental para o progreso da ciencia, xa que fai posible a comprensión dos resultados de cada experimento, posibilitando a detección de erros e a mellora das técnicas científicas.

Capítulo 1

Introdución biolóxica

Neste capítulo introducimos o marco biolóxico dos problemas que estudaremos nos seguintes capítulos, co obxectivo de contextualizar ditos problemas e coñecer o material co que imos a traballar.

O **ADN**, siglas de ácido desoxirribonucleico, é o noso material xenético e está composto dunha longa secuencia de nucleótidos. Tal e como se mostra na Figura 1.1, cada **nucleótido** está composto por un grupo fosfato, un monosacárido (azucre de cinco carbonos) e unha base nitroxenada.

O punto máis importante da composición do ADN para a análise das secuencias é que o ADN conta con catro tipos de bases nitroxenadas: adenina (A), guanina (G), citosina (C) e timina (T). Podemos ver as diferentes bases na chave da Figura 1.1.

A análise das secuencias de ADN centra o seu estudo principalmente nestas catro bases nitroxenadas, xa que son as que determinan a información xenética. Como podemos ver, o elemento diferenciador do nucleótido é a base nitroxenada. Polo que, de aquí en adiante cando fagamos referencia aos nucleótidos referímonos ás bases nitroxenadas. Debido a súa relevancia, estúdanse as posicións de cada tipo de nucleótido nunha secuencia de ADN.

As **secuencias de ADN** son fragmentos de ADN que representaremos polos seus nucleótidos tal como indicamos antes. Existen numerosas técnicas de secuenciación do ADN que van evolucionando cos avances científicos.

Á hora de obter as secuencias, unha forma é unir fragmentos de ADN e formar contigs, secuencias máis longas formadas polo solapamento destes fragmentos. Este proceso recibe o nome de **ensamblado**.

A estrutura tridimensional do ADN consiste nunha dobre hélice orientada de maneira que se

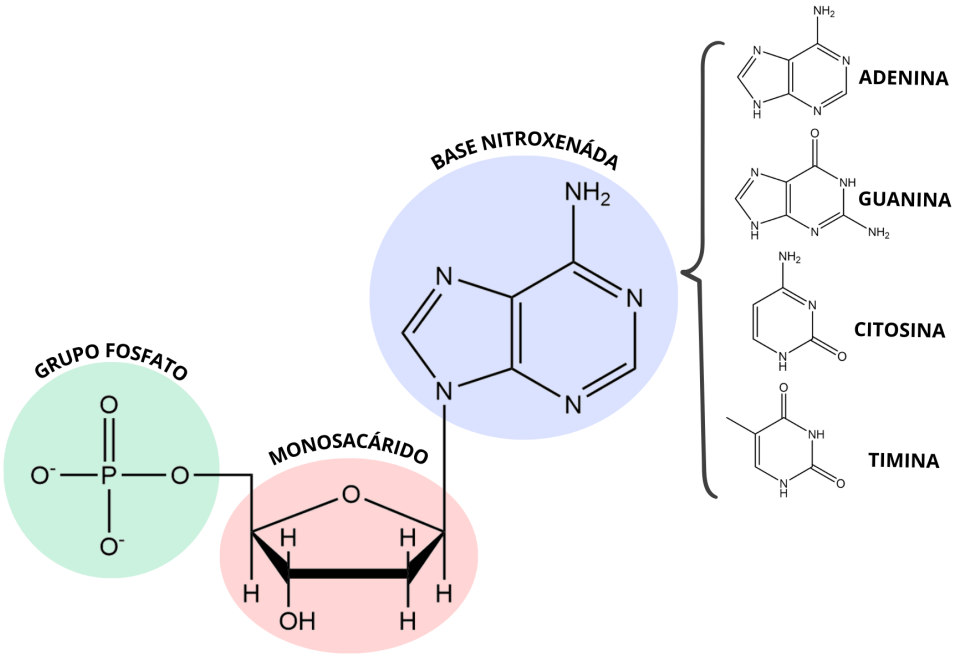

Figura 1.1: Estrutura dun nucleótido xerada con ChemDraw.

forman enlaces entre bases de cadeas opostas, apareándose a adenina coa timina, e a guanina coa citosina. Estes enlaces manteñen unidas as dúas cadeas do ADN que forman a dobre hélice. Debido ao apareamento específico as dúas cadeas de ADN son complementarias, cada cadea contén a información necesaria para determinar a secuencia de nucleótidos da outra cadea.

O aliñamento de secuencias de ADN é unha técnica empregada para comparar dúas ou máis secuencias de ADN de distinta procedencia. Trátase de buscar as súas similitudes que poderían indicar que existe algunha relación evolutiva (que teñen algún antepasado en común) ou funcional (cumpren funcións similares no organismo).

A análise de secuencias de ADN ten unha ampla base matemática. Consideraremos unha secuencia de ADN como unha lista de obxectos ou eventos formados por elementos que poden ocupar aleatoriamente diferentes posicións na secuencia.

O **xenoma** é o conxunto de todo o material xenético dun organismo. O noso material xenético está contido en 23 pares de longas moléculas de ADN chamadas cromosomas e consta de aproximadamente tres mil millóns de nucleótidos.

Neste traballo empregaremos un xenoma para ilustrar un caso práctico. Debemos ter en conta que o tamaño e organización do xenoma non é o mesmo para organismos formados por células eucariotas (como o ser humano), que para organismos formados por células procariotas.

As células divídense en dúas clases principais segundo se posúen ou non unha estrutura específica para almacenar o seu material xenético. Esta estrutura chámase núcleo. Así, distinguimos as **células eucariotas**, que presentan núcleo que separa o ADN do resto do contido celular, e as **células procariotas**, que carecen de núcleo.

No obstante, que as células procariotas non posúan núcleo non quere dicir que non teñan material xenético. De feito o seu material xenético está formado por unha soa molécula de ADN en forma circular, a diferencia das eucariotas que está formado por múltiples moléculas lineais de ADN.

Ademais o xenoma das células procariotas é menos complexo ca o das células eucariotas. Mentres que xenoma das procariotas consta de entre un e cinco millóns de nucleótidos o das eucariotas e moito máis grande podendo chegar a cinco mil millóns de nucleótidos.

As **bacterias** son microorganismos unicelulares procariotas. Estes organismos poden presentar diversas formas, o que permite clasificalos segundo esta característica en: cocos (forma esférica), bacilos (forma de bastón) e espirilos (forma de espiral). As bacterias están presentes en todo o noso entorno, xa que viven nunha ampla variedade de ambientes como terra, auga e o interior doutros organismos.

A presenza das bacterias en todos estes ambientes, sumado ao feito de que o seu xenoma é pequeno en comparación co xenoma doutros seres vivos, convérteas en candidatas perfectas para a análise de secuencias de ADN.

Os conceptos ilustrados neste capítulo están recollidos en [2].

Capítulo 2

Probabilidades e variables aleatorias

Neste capítulo centrarémonos nos conceptos da teoría de probabilidade necesarios para abordar a análise de secuencias de ADN. É necesario manexar probabilidades e entender as variables aleatorias e as súas distribucións debido a súa importancia no campo da bioinformática. Os contidos recollidos neste capítulo proveñen de [8], [4] e [5].

2.1. Probabilidades

Un **experimento aleatorio** é aquel no que non sabemos con seguridade cal de todos os seus posibles resultados vai ocorrer cando o executamos, como lanzar un dado ou sacar unha carta da baralla. O conxunto de todos os resultados posibles dun experimento aleatorio é o **espazo mostral** e denótase como Ω. No caso do dado, o espazo mostral é o conxunto $\{1, 2, 3, 4, 5, 6\}$ e no caso das cartas, o espazo mostral é o conxunto dos números enteiros que vai dende 1 a 40, se se emprega a baralla española.

Consideramos agora a familia de subconxuntos de Ω, $\mathcal{F}$, cumprindo os seguintes requisitos:

(i) Se $A \in \mathcal{F}$ o seu complementario, $A^c = \Omega - A$, tamén pertence a $\mathcal{F}$.

(ii) Se $\{A_n\}$ é unha colección finita ou numerable de conxuntos de $\mathcal{F}$, a súa unión, $\bigcup_n A_n$, está en $\mathcal{F}$.

Estas dúas propiedades definen a $\mathcal{F}$ coma unha σ-álxebra de conxuntos de Ω onde cada elemento de $\mathcal{F}$ denomínase **suceso**, polo que a familia $\mathcal{F}$ recibe o nome de **σ-álxebra de sucesos**. Ademais, chamamos **suceso elemental** a cada elemento de Ω.

Se Ω é finito, traballaremos con $\mathcal{F} = \mathcal{P}(\Omega)$, onde $\mathcal{P}(\Omega)$ é a colección de todos os subconxuntos de Ω. En xeral, traballaremos con $\Omega = \mathbb{R}$ e escollemos como σ-álxebra de sucesos a *σ-álxebra de*

Borel en $\mathbb{R}$. A σ-álxebra de Borel en $\mathbb{R}$, denotada por $\mathcal{B}(\mathbb{R})$, está xerada por unións e intersecións numerables dos intervalos $(-\infty, x]$ con $x \in \mathbb{R}$ e os seus elementos denomínanse conxuntos de Borel.

Á hora de executar un experimento aleatorio non sabemos que é o que vai ocorrer, polo que nos interesa medir dalgunha forma a incerteza dos posibles resultados que podemos obter. Unha forma de medir esta incerteza é a seguinte.

Consiste en repetir un número suficientemente grande de veces un experimento aleatorio e dar como medida a frecuencia relativa dun suceso A (o número de veces que obtemos o resultado A entre o número de repeticións n) que se vai achegando a un determinado valor. Desta forma, se puidésemos executar o experimento un número infinito de veces chegaríamos a un valor que podemos intuír como a probabilidade do suceso A. Isto baséase na **lei da estabilidade das frecuencias**.

Por exemplo, se lanzamos un dado podemos observar que a proporción de veces que sacamos un catro, é dicir, a frecuencia relativa deste resultado, vaise achegando a un certo valor a medida que repetimos o lanzamento. No obstante, quedarémonos só coa idea intuitiva, xa que non empregaremos dito enfoque da probabilidade.

A continuación presentamos a definición de probabilidade coa que traballaremos ao longo do documento.

Definición 2.1. (de Kolmogorov) Dado un par $(\Omega, \mathcal{F})$, unha **probabilidade** é unha aplicación $P : \mathcal{F} \to \mathbb{R}$ que verifica

(i) $P(A) \geq 0$ para todo $A \in \mathcal{F}$.

(ii) Para calquera colección numerable de sucesos $A_n \subset \mathcal{F}$, disxuntos entre si, cúmprese

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n).$$

(iii) $P(\Omega) = 1$.

Se P é unha probabilidade en $(\Omega, \mathcal{F})$, denomínase **espazo de probabilidade** á terna $(\Omega, \mathcal{F}, P)$ e o valor $P(A)$ asociado a cada suceso $A \in \mathcal{F}$ recibe o nome de probabilidade de A.

Chamamos **suceso imposible** a un suceso A tal que $P(A) = 0$ e **suceso seguro** a un suceso A tal que $P(A) = 1$.

A continuación recóllense as principais propiedades da probabilidade derivadas da definición anterior.

Proposición 2.2. *Consideremos os sucesos A e B. Verifícanse as seguintes propiedades:*

(i) $P(A^c) = 1 - P(A)$.

(ii) $P(\varnothing) = 0$.

(iii) *Se* $B \subset A$, *entón* $P(B) \leq P(A)$.

(iv) *(Principio de Inclusión-Exclusión) Dados n sucesos* $A_1, A_2, ..., A_n \in \mathcal{F}$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k)$$

$$+ \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

Presentamos a regra de Laplace que nos permite calcular a probabilidade dun suceso calquera, baixo certas condicións que veremos a continuación.

Dado un experimento aleatorio con espazo mostral finito Ω no que todos os sucesos elementais teñen a mesma probabilidade, a **Regra de Laplace** permítenos calcular a probabilidade dun suceso calquera A contando o número de casos da seguinte forma

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{nº de casos favorables ao suceso A}}{\text{nº de casos posibles}}.$$

A continuación exemplificamos un problema de análise de secuencias relacionado coas probabilidades.

Exemplo 2.3. Imaxinemos que tomamos dúas secuencias aleatorias de ADN, cada unha delas composta por tres nucleótidos. Recordemos que consideramos unha secuencia como unha lista de obxectos ou eventos formados por elementos (as letras A, C, G e T), que poden ocupar diferentes posicións na secuencia.

Calculemos entón a probabilidade de que estas secuencias teñan polo menos unha coincidencia. Chamamos coincidencia ao feito de ter o mesmo nucleótido nunha mesma posición, por exemplo, que ambas teñan como primeiro nucleótido a adenina.

Para calcular a probabilidade de que exista polo menos unha coincidencia empregaremos a propiedade (i) e centrarémonos no cálculo da probabilidade do suceso complementario, é dicir, que as secuencias non coincidan en ningunha posición.

Comezamos contando os casos posibles, todas as posibles parellas de secuencias de lonxitude tres nucleótidos. Notemos que nas secuencias importa a orde dos elementos e ademais estes poden repetirse. É dicir, as posibles secuencias son variacións con repeticións de orde 3 dos 4 elementos, polo que hai $4^3 = 64$ posibles variacións de cada secuencia. Como estamos estudando dúas secuencias, seguindo o mesmo razoamento, o número de posibles parellas é $64^2 = 4096$. Existen 4096 casos favorables.

Agora ben, destes 4096 casos, precisamos contar en cantos as dúas secuencias non coinciden en ningunha posición. Escollemos só unha das 64 posibles secuencias, por exemplo AGC. A outra secuencia ten que ter unha letra distinta da primeira en cada posición. Logo, a outra secuencia ten 3 posibilidades para a primeira posición (G, C e T), 3 posibilidades para a segunda posición (A, C e T) e outras 3 para a terceira posición (A, G e T). Aplicando o *principio da multiplicación*[1] obtemos 27 posibles secuencias con letras distintas a AGC en cada posición. Pero como en realidade podemos escoller a secuencia de referencia de 64 formas diferentes, os casos favorables aumentan a $27 \cdot 64 = 1728$.

Finalmente, empregando a regra de Laplace, a probabilidade de que non exista ningunha coincidencia é $\frac{1728}{4096} \approx 0,422$. Polo tanto, a probabilidade de que exista unha ou máis coincidencias é aproximadamente $1 - 0,422 = 0,578$.

Nalgúns casos a probabilidade de que ocorra un suceso depende de se outro suceso ocorreu ou non. Por exemplo, supoñamos que lanzamos un dado e queremos calcular a probabilidade de que "o resultado sexa par", suceso A, mais sabemos que o número que saíu "é maior que tres", suceso B. Polo tanto, a probabilidade de obter A sabendo que ocorreu B non é a mesma que se non coñecésemos a información dada polo suceso B. Nese caso, temos a probabilidade condicionada de A dado B.

Definición 2.4. Se $B \in \mathcal{F}$ é un suceso con $P(B) > 0$, para cada $A \in \mathcal{F}$, denomínase **probabilidade de** A **condicionada por** B a

$$P(A \mid B) = \frac{P(A \cup B)}{P(B)}.$$

A continuación, presentamos tres resultados básicos para o cálculo de probabilidades.

Proposición 2.5. *(Regra do produto) Dados os sucesos $A_1, A_2, \dots, A_n$ tales que $P\left(\bigcap_{i=1}^{n-1} A_i\right) > 0$. Entón*

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdot \cdots \cdot P\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right).$$

Proposición 2.6. *(Fórmula das probabilidades totais) Dados os sucesos $B_1, B_2, \dots, B_n$ tales que*

[1] *Principio da multiplicación* [1]. Se unha tarefa está formada por n subtarefas sucesivas e independentes e cada un dos conxuntos A_i representa as distintas formas de realizar a subtarefa i, entón o produto cartesiano $A_1 \times A_2 \times \cdots \times A_n$ representa as distintas formas de realizar a tarefa principal e o principio da multiplicación dinos que, sendo $A_1, A_2, \dots, A_n$ conxuntos finitos, tense

$$|A_1 \times A_2 \times \cdots \times A_n| = |A_1| \times |A_2| \times \cdots \times |A_n|,$$

onde $|A_i|$ representa o cardinal de A_i, é dicir, o número de elementos do conxunto.

$\bigcup_{i=1}^{n} B_i = \Omega$ *e* $B_i \cap B_j = \emptyset$ *para* $i \neq j$*, entón para calquera suceso* A *cúmprese*

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i).$$

Proposición 2.7. *(Fórmula de Bayes) Se* A *é un suceso e consideramos os sucesos* $B_1, B_2, \ldots, B_n$ *tales que* $\bigcup_{i=1}^{n} B_i = \Omega$ *e* $B_i \cap B_j = \emptyset$*, para* $i \neq j$*, entón*

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)}, \quad \forall i \in \{1, 2, \ldots, n\}.$$

2.2. Variables aleatorias unidimensionais

Consideramos o experimento aleatorio de lanzar dúas moedas. Os posibles resultados serían: que saísen dúas cruces (0 caras), que saíse cara e cruz (1 cara) ou que saísen dúas caras (2 caras). A variable que describe o número de caras resultantes do lanzamento de dúas moedas tomando os valores 0, 1 ou 2, é a que chamaremos variable aleatoria. A continuación presentamos a definición formal de variable aleatoria.

Definición 2.8. Denomínase **variable aleatoria** definida nun espazo de probabilidade $(\Omega, \mathcal{F}, P)$ a calquera función $X : \Omega \to \mathbb{R}$ tal que

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad \text{para cada conxunto } B \in \mathcal{B}(\mathbb{R}).$$

Como calquera boreliano $B \in \mathcal{B}(\mathbb{R})$ pode obterse a partir dos intervalos $(-\infty, x]$ mediante unións e interseccióǹs numerables, a definición anterior é equivalente a que X é unha variable aleatoria se se verifica

$$X^{-1}\left((-\infty, x]\right) = \{\omega \in \Omega : X(\omega) \leq x\} = \{X \leq x\} \in \mathcal{F}, \quad \forall x \in \mathbb{R}.$$

Como consecuencia, se X é unha variable aleatoria tamén os conxuntos $\{X = a\}, \{a \leq X < b\}, \{X > b\} \in \mathcal{F}$ para todo $a, b \in \mathbb{R}$.

Teorema 2.9. *A variable aleatoria* X *definida nun espazo de probabilidade* $(\Omega, \mathcal{F}, P)$ *induce un espazo de probabilidade* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ *mediante a correspondencia*

$$P_X(B) = P\left\{X^{-1}(B)\right\} = P\left\{\omega \in \Omega : X(\omega) \in B\right\}, \quad \textit{para cada } B \in \mathcal{B}(\mathbb{R}),$$

onde P_X *recibe o nome de* ***distribución de probabilidade da variable aleatoria*** X*.*

As funcións de distribución asociadas a variables aleatorias, permiten coñecer a distribución de probabilidade de dita variable aleatoria. Primeiro introducimos o concepto xeral de función de distribución.

Definición 2.10. Unha función $F : \mathbb{R} \to [0, 1]$ que verifica a seguintes propiedades

(i) F é non decrecente, é dicir, $F(x) \leq F(y)$ sempre que sexa $x < y$.
(ii) F é continua pola dereita, isto é, para cada $x \in \mathbb{R}$, cúmprese

$$\lim_{h \to 0} F(x + h) = F(x).$$

(iii) $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to \infty} F(x) = 1$,

recibe o nome de **función de distribución**.

Definición 2.11. Sexa X unha variable aleatoria definida no espazo de probabilidade $(\Omega, \mathcal{F}, P)$ e P_X a distribución de probabilidade da variable aleatoria X. A función $F : \mathbb{R} \to [0, 1]$ definida como

$$F(x) = P_X((-\infty, x]) = P\{X^{-1}(-\infty, x]\} = P\{\omega \in \Omega : X(\omega) \leq x\},$$

recibe o nome de **función de distribución da variable aleatoria** X.
É habitual escribir

$$F(x) = P\{X \leq x\}.$$

Estudaremos distintos tipos de variables aleatorias. A continuación presentamos as súas definicións.

Definición 2.12. Unha variable aleatoria X definida no espazo de probabilidade $(\Omega, \mathcal{F}, P)$ dise que é unha **variable aleatoria discreta** se existe un conxunto contable $E \subseteq \mathbb{R}$ tal que $P\{X \in E\} = 1$. A colección de números p_i que satisfai que $P\{X = x_i\} = p_i \geq 0$, para todo i e $\sum_{i=1}^{\infty} p_i = 1$, recibe o nome de **función de masa de probabilidade** da variable aleatoria X.

Polo tanto súa función de distribución é

$$F(x) = \sum_{x_i \leq x} P\{X = x_i\}.$$

Vemos que, neste caso, a función de distribución ten descontinuidades de salto en cada un dos puntos nos que toma valores a variable aleatoria.

O número que sae ao lanzar un dado ou o resultado de lanzar unha moeda son exemplos de variables aleatorias discretas, xa que toman valores nun conxunto finito.

Definición 2.13. Unha variable aleatoria X definida no espazo de probabilidade $(\Omega, \mathcal{F}, P)$ con función de distribución $F : \mathbb{R} \to [0, 1]$, dise que é unha **variable aleatoria continua** se existe unha función non negativa $f : \mathbb{R} \to \mathbb{R}$ tal que

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \quad \text{para cada } x \in \mathbb{R}.$$

En tal caso, f recibe o nome de **función de densidade** da variable aleatoria X.

Neste caso a función de distribución é continua en todo punto.

A función de densidade f verifica que

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Ademais, en todo punto $x \in \mathbb{R}$ no que f sexa continua, $F'(x) = f(x)$.

Notemos que se X é unha variable aleatoria continua, dados $a, b \in \mathbb{R}$ tal que $a < b$, entón

$$P\{a < X \leq b\} = \int_{-\infty}^{b} f(x)\, dx - \int_{-\infty}^{a} f(x)\, dx = \int_{a}^{b} f(x)\, dx.$$

O tempo de espera dunha cola, o peso dunha persoa ou a distancia recorrida por un vehículo son exemplos de variables aleatorias continuas, xa que toman valores nun intervalo de $\mathbb{R}$.

A continuación introducimos o concepto de distribución truncada que empregaremos no capítulo de ensamblado.

Definición 2.14. Sexa X unha variable aleatoria continua, definida no espazo de probabilidade $(\Omega, \mathcal{F}, P)$, con función de distribución $f : \mathbb{R} \to \mathbb{R}$, e $T \in \mathcal{B}(\mathbb{R})$ tal que $0 < P\{X \in T\} < 1$. Chamamos **función de distribución truncada de** X á función de distribución definida como

$$P\{X \leq x \mid X \in T\} = \frac{P\{X \leq x, X \in T\}}{P\{X \in T\}} = \frac{\int_{(-\infty, x] \cap T} f(y)\ dy}{\int_T f(y)\ dy}, \quad \forall x \in \mathbb{R}.$$

A función de densidade da distribución truncada vén dada por

$$h(x) = \begin{cases} \frac{f(x)}{\int_T f(y)\ dy,} & x \in T, \\ 0 & x \notin T. \end{cases} \tag{2.1}$$

A media ou esperanza dunha variable aleatoria é unha medida de localización que nos dá información sobre a localización dos valores que toma a variable aleatoria.

Definición 2.15. Se X é unha variable aleatoria discreta dicimos que existe a **media** ou esperanza da variable aleatoria X se

$$\sum_{i=1}^{\infty} |x_i| P\{X = x_i\} < \infty.$$

En tal caso vén dada por

$$E[X] = \sum_{i=1}^{\infty} x_i P\{X = x_i\}.$$

Por outro lado, se X é unha variable aleatoria continua, con función de densidade f, dicimos que existe a **media** ou esperanza da variable aleatoria X se

$$\int_0^{\infty} |x| f(x)\, dx < \infty.$$

En tal caso vén dada por

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

A continuación, recóllense algunhas propiedades da media.

Proposición 2.16.

(i) Sexan X_1, X_2 dúas variables aleatorias. Se $E[X_1]$ e $E[X_2]$ son finitas e $a, b \in \mathbb{R}$, logo

$$E[aX_1 + bX_2] = aE[X_1] + bE[X_2].$$

(ii) Se $X_1, X_2, \dots, X_n$ son variables aleatorias independentes, cúmprese

$$E[X_1 \cdot X_2 \cdot \cdots \cdot X_n] = E[X_1]E[X_2] \cdot \cdots \cdot E[X_n].$$

Demostración.

(i) Supoñamos que X_1 e X_2 son variables aleatorias continuas. Sexa $f : \mathbb{R}^2 \to \mathbb{R}$ a función de densidade conxunta de X_1 e X_2, e $f_i : \mathbb{R} \to \mathbb{R}$ a función de densidade marxinal de X_i para $i \in \{1, 2\}$, entón

$$\begin{aligned} E[aX_1 + bX_2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax_1 + bx_2) f(x_1, x_2)\ dx_1\ dx_2 \\ &= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)\ dx_2\ dx_1 + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)\ dx_1\ dx_2 \\ &= a\int_{-\infty}^{\infty} x_1 f_1(x_1)\ dx_1 + b\int_{-\infty}^{-\infty} x_2 f_2(x_2)\ dx_2 \\ &= aE[X_1] + bE[X_2]. \end{aligned}$$

Vexamos que tamén se cumpre no caso de que X_1 e X_2 sexan dúas variables aleatorias

discretas.

$$\begin{aligned}
E[aX_1 + bX_2] &= \sum_{x_1,x_2} (ax_1 + bx_2) P\{X_1 = x_1, X_2 = x_2\} \\
&= \sum_{x_1} ax_1 \sum_{x_2} P\{X_1 = x_1, X_2 = x_2\} + \sum_{x_2} bx_2 \sum_{x_1} P\{X_1 = x_1, X_2 = x_2\} \\
&= \sum_{x_1} ax_1 P\{X_1 = x_1\} + \sum_{x_2} bx_2 P\{X_2 = x_2\} \\
&= aE[X_1] + bE[X_2].
\end{aligned}$$

(ii) Supoñamos que son continuas. Pola independencia das variables aleatorias $X_1, X_2, \dots, X_n$ a súa función de densidade conxunta $f : \mathbb{R}^n \to \mathbb{R}$ verifica $f(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \cdot \cdots \cdot f_n(x_n)$, onde para cada $i \in \{1, 2, \dots, n\}$ $f_i : \mathbb{R} \to \mathbb{R}$ é a función de densidade de X_i, logo

$$\begin{aligned}
E[X_1 \cdot X_2 \cdot \cdots \cdot X_n] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \cdot \int_{-\infty}^{\infty} x_1 x_2 \cdot \cdots \cdot x_n f(x_1, x_2, \dots, x_n)\ dx_1 dx_2 \cdot \cdots \cdot dx_n \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \cdot \int_{-\infty}^{\infty} x_1 x_2 \cdot \cdots \cdot x_n f_1(x_1) f_2(x_2) \cdot \cdots \cdot f_n(x_n)\ dx_1 dx_2 \cdot \cdots \cdot dx_n \\
&= \int_{-\infty}^{\infty} x_1 f_1(x_1)\ dx_1 \int_{-\infty}^{\infty} x_2 f_2(x_2)\ dx_2 \cdot \cdots \cdot \int_{-\infty}^{\infty} x_n f_n(x_n)\ dx_n \\
&= E[X_1] E[X_2] \cdot \cdots \cdot E[X_n].
\end{aligned}$$

Vexamos que tamén se cumpre no caso de que sexan discretas. Pola independencia das variables aleatorias $X_1, X_2, \dots, X_n$ a súa función de masa de probabilidade verifica $P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = P\{X_1 = x_1\} P\{X_2 = x_2\} \cdot \cdots \cdot P\{X_n = x_n\}$, entón

$$\begin{aligned}
E[X_1 \cdot X_2 \cdot \cdots \cdot X_n] &= \sum_{x_1,x_2,\dots,x_n} x_1 x_2 \cdot \cdots \cdot x_n P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} \\
&= \sum_{x_1,x_2,\dots,x_n} x_1 x_2 \cdot \cdots \cdot x_n P\{X_1 = x_1\} P\{X_2 = x_2\} \cdot \cdots \cdot P\{X_n = x_n\} \\
&= \left(\sum_{x_1} x_1 P\{X_1 = x_1\} \right) \left(\sum_{x_2} x_2 P\{X_1 = x_2\} \right) \cdot \cdots \cdot \left(\sum_{x_n} x_n P\{X_n = x_n\} \right) \\
&= E[X_1] E[X_2] \cdot \cdots \cdot E[X_n].
\end{aligned}$$

□

2.3. Vectores aleatorios e as súas distribucións

Consideramos o experimento aleatorio de lanzar dous dados, un azul e outro vermello. Queremos estudar se existe algunha relación entre o resultado do dado vermello e a suma dos resultados

de ambos dados. Desta forma traballaremos con dúas variables aleatorias, X_1, que describe o número que obtemos no dado vermello, e X_2, que describe a suma do resultado do dado vermello e do dado azul. Estas variables forman o que chamaremos vector aleatorio.

Definición 2.17. A unha colección de n variables aleatorias $X_1, X_2, \dots, X_n$, definida sobre o mesmo espazo de probabilidade $(\Omega, \mathcal{F}, P)$, chamarémola **vector aleatorio** de dimensión n

$$\mathbf{X} = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n,$$

onde $\mathbf{X} : \Omega \to \mathbb{R}^n$, é unha variable aleatoria. É dicir, para cada conxunto I da forma

$$I = \{(x_1, x_2, \dots, x_n) : -\infty < x_i \leq y_i, y_i \in \mathbb{R}, i = 1, 2, \dots, n\},$$

verifícase que

$$\mathbf{X}^{-1}(I) = \{\omega : X_1(\omega) \leq y_1, X_2(\omega) \leq y_2, \dots, X_n(\omega) \leq y_n\} \in \mathcal{F}, \quad \text{para cada } y_i \in \mathbb{R}.$$

A función de distribución asociada a unha variable aleatoria permite coñecer a súa distribución de probabilidade. A continuación presentamos a definición de función de distribución da variable aleatoria $\mathbf{X}$.

Definición 2.18. Dado un vector aleatorio $\mathbf{X} = (X_1, X_2, \dots, X_n)$ a función $F : \mathbb{R}^n \to [0, 1]$ definida como

$$F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\}, \quad \forall (x_1, x_2, \dots, x_n) \in \mathbb{R}^n,$$

é a función de distribución de $\mathbf{X}$ tamén chamada **función de distribución conxunta** de $X_1, X_2, \dots, X_n$.

A función de distribución conxunta, por ser función de distribución da variable aleatoria $\mathbf{X}$, verifica que é non decrecente e continua pola dereita e ademais

- $\lim_{x_i \to -\infty} F(x_1, \dots, x_n) = 0, \quad \forall (x_1, \dots, x_{i-1}, x_{i+1} \dots, x_n) \in \mathbb{R}^{n-1}$.
- $\lim_{x_1 \to \infty, \dots, x_i \to \infty, \dots, x_n \to \infty} F(x_1, \dots, x_n) = 1$.

Estudaremos distintos tipos de vectores aleatorios. A continuación presentamos as súas definicións.

Definición 2.19. Un vector aleatorio $\mathbf{X} = (X_1, X_2, \dots, X_n)$ dise que é un **vector aleatorio discreto** se $X_1, X_2, \dots, X_n$, son variables aleatorias discretas. A **función de masa de probabilidade conxunta** do vector aleatorio $(X_1, X_2, \dots, X_n)$ vén dada por

$$P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}, \quad \forall (x_1, x_2, \dots, x_n) \in \mathbb{R}^n,$$

satisfacendo

$$P\{X_1 = x_1, \dots, X_n = x_n\} \geq 0, \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \text{e} \quad \sum_{x_1, \dots, x_n} P\{X_1 = x_1, \dots, X_n = x_n\} = 1.$$

Definición 2.20. Un vector aleatorio $\mathbf{X} = (X_1, X_2, \dots, X_n)$ con función de distribución conxunta $F : \mathbb{R}^n \to [0,1]$, dise que é un **vector aleatorio continuo** se existe unha función non negativa $f : \mathbb{R}^n \to \mathbb{R}$ tal que

$$F(y_1, y_2, \dots, y_n) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \cdots \cdot \int_{-\infty}^{y_n} f(x_1, x_2, \dots, x_n)\ dx_1 dx_2 \cdot \cdots \cdot dx_n, \quad \forall (y_1, y_2, \dots, y_n) \in \mathbb{R}^n.$$

En tal caso, f recibe o nome de **función de densidade conxunta**.

De forma análoga ao caso unidimensional, a función de densidade conxunta verifica

$$\int_{-\infty}^{\infty} \cdots \cdot \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\ dx_1 \cdot \cdots \cdot dx_n = 1.$$

Ademais, se f é continua en $(y_1, \dots, y_n) \in \mathbb{R}^n$, logo

$$\frac{\partial^n F(y_1, \dots, y_n)}{\partial y_1 \cdot \cdots \cdot \partial y_n} = f(y_1, \dots, y_n).$$

É claro que a altura dunha persoa mantén certa relación coa lonxitude do seu fémur, xa que é unha parte da altura total. Pola contra, a altura non está relacionada co nivel de colesterol. Por iso dicimos que variable aleatoria que describe a altura e a variable aleatoria que describe o nivel de colesterol son independentes. A continuación damos a definición formal de independencia de variables aleatorias.

Definición 2.21. As variables aleatorias $X_1, X_2, ..., X_n$, definidas no espazo de probabilidade $(\Omega, \mathcal{F}, P)$, son independentes se, e só se,

$$\begin{aligned} P\{X_1 \le x_1, X_2 \le x_2, ..., X_n \le x_n\} &= P\{X_1 \le x_1\}\, P\{X_2 \le x_2\} \cdot \cdots \cdot P\{X_n \le x_n\} \\ &= F_1(x_1) F_2(x_2) \cdot \cdots \cdot F_n(x_n), \end{aligned}$$

calquera que sexan $x_1, x_2, ..., x_n \in \mathbb{R}$, onde $F_1, F_2, ..., F_n$ son as respectivas funcións de distribución.

Polo tanto, no caso continuo a súa función de densidade conxunta $f : \mathbb{R}^n \to \mathbb{R}$ verifica

$$f(x_1, \dots, x_n) = f_1(x_1) \cdot \cdots \cdot f_n(x_n), \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n,$$

onde $f_i : \mathbb{R} \to \mathbb{R}$ é a función de densidade de X_i, para cada $i \in \{1, \dots, n\}$. Analogamente, no caso discreto a súa función de masa de probabilidade verifica

$$P\{X_1 = x_1, \dots, X_n = x_n\} = P\{X_1 = x_1\} \cdot \cdots \cdot P\{X_n = x_n\}, \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Consideremos un vector aleatorio $(X_1, X_2, \dots, X_n)$. Imaxinemos que queremos coñecer a distribución de probabilidade dunha variable aleatoria X_i, para certo $i \in \{1, \dots, n\}$, ignorando os efectos do resto de variables aleatorias $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. Esta distribución de probabilidade recibe o nome de distribución marxinal de X_i, e non é máis que a distribución de probabilidade da variable aleatoria unidimensional X_i.

Definición 2.22. Dado un vector aleatorio discreto $\mathbf{X} = (X_1, \dots, X_n)$ a **función de masa de probabilidade marxinal** de X_i vén dada por

$$P\{X_i = x_i\} = \sum_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n} P\{X_1 = x_1, \dots, X_n = x_n\}, \quad \forall x_i \in \mathbb{R},$$

verificando

$$P\{X_i = x_i\} \geq 0, \ \forall x_i \in \mathbb{R}, \quad \text{e} \quad \sum_{x_i} P\{X_i = x_i\} = 1.$$

Definición 2.23. Sexa $\mathbf{X} = (X_1, \dots, X_n)$ un vector aleatorio continuo e f a súa función de densidade conxunta, a función $f_i : \mathbb{R} \to \mathbb{R}$ dada por

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \cdot \int_{-\infty}^{\infty} f(x_1, \dots, x_n) \ dx_1 \cdot \cdots \cdot dx_{i-1} dx_{i+1} \cdot \cdots \cdot dx_n,$$

que verifica

$$f_i(x_i) \geq 0 \quad \text{e} \quad \int_{-\infty}^{\infty} \cdots \cdot \int_{-\infty}^{\infty} f_i(x_i) \ dx_i = 1,$$

é a **función de densidade marxinal** de X_i.

Analogamente á probabilidade condicionada xorden as distribucións condicionadas das variables aleatorias. A continuación presentamos algunhas distribucións condicionadas dadas dúas variables aleatorias.

Definición 2.24. Dado $\mathbf{X} = (X_1, X_2)$ un vector aleatorio discreto tal que $P\{X_2 = x_2\} > 0$, defínese a **función de masa de probabilidade de X_1 condicionada a que $X_2 = x_2$** como

$$P\{X_1 = x_1 \mid X_2 = x_2\} = \frac{P\{X_1 = x_2, X_2 = x_2\}}{P\{X_2 = x_2\}},$$

verificando

$$P\{X_1 = x_1 \mid X_2 = x_2\} \geq 0, \quad \forall x_1 \in \mathbb{R}, \quad \text{e} \quad \sum_{x_1} P\{X_1 = x_1 \mid X_2 = x_2\} = 1.$$

Nótese que, se as variables aleatorias X_1 e X_2 son independentes, entón

$$P\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1\} P\{X_2 \ = x_2\},$$

e así,

$$P\{X_1 = x_1 \mid X_2 = x_2\} = \frac{P\{X_1 = x_1, X_2 = x_2\}}{P\{X_2 = x_2\}} = P\{X_1 = x_1\}.$$

Definición 2.25. Sexa $\mathbf{X} = (X_1, X_2)$ un vector aleatorio continuo con función de densidade conxunta $f : \mathbb{R}^2 \to \mathbb{R}$ e $f_2 : \mathbb{R} \to \mathbb{R}$ a función de densidade de X_2. En todo punto (x_1, x_2) onde f é continua e $f_2(x_2) > 0$ e continua, a **función de densidade de X_1 condicionada a X_2** existe e vén dada por

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)},$$

verificando

$$f_{X_1|X_2}(x_1 \mid x_2) \geq 0 \quad \text{e} \quad \int_{-\infty}^{\infty} f_{X_1|X_2}(x_1 \mid x_2) \ dx_1 = 1.$$

2.4. Algunhas distribucións de variables aleatorias unidimensionais

Nesta sección recóllense as distribucións de probabilidade que empregaremos en capítulos seguintes xunto coa súa media.

Supoñamos que temos un experimento aleatorio e interésanos estudar se ocorre ou non un determinado suceso A. Neste caso, se ocorre A diremos que hai un "éxito" e cada vez que realicemos o experimento é un "ensaio".

A **distribución binomial** é unha distribución discreta que describe o número de éxitos en n ensaios dun experimento aleatorio, verificando certas condicións. A probabilidade de éxito p ten que ser a mesma en todos os ensaios, o número de ensaios n é un número fixo e os ensaios deben ser independentes.

Se X é unha variable aleatoria que segue unha distribución binomial de parámetros n e p, $X \sim Bin(n,p)$, X é o número de éxitos total en n ensaios e toma valores no conxunto $\{0,1,2...,n\}$.

Supoñamos que lanzamos unha moeda 10 veces e consideramos como un "éxito" que saia cara, logo a variable aleatoria que describe o número de caras ao tirar unha moeda 10 veces segue unha distribución binomial de parámetro $p = \frac{1}{2}$ e $n = 10$.

A función de masa de probabilidade de X describe a probabilidade de obter i éxitos en n intentos e vén dada por

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0,1,2,...,n, \quad \text{onde} \quad \binom{n}{i} = \frac{n!}{i!(n-i)!}.$$

Supoñamos que $X \sim Bin(n,p)$ e calculemos a súa media.

$$\begin{aligned} E[X] &= \sum_{i=0}^{n} iP\{X=i\} = \sum_{i=1}^{n} i\binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!} p^i (1-p)^{n-i} \\ &= \sum_{k=0}^{n-1} \frac{n!}{k!(n-k-1)!} p^{k+1} (1-p)^{n-k-1} = np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k} \\ &= np(p+1-p)^{n-1} = np, \end{aligned}$$

onde na terceira igualdade reindexamos tomando $k = i-1$, polo tanto

$$E[X] = np \tag{2.2}$$

Supoñamos que estamos interesados en un suceso A (éxito) ao realizar un experimento aleatorio.

A **distribución xeométrica** é unha distribución discreta que describe o número de éxitos antes de obter o primeiro fracaso dun experimento aleatorio que se executa repetidas veces ata que se

chega ao primeiro fracaso. Neste caso, o numero de ensaios non é fixo, finaliza cando se atopa o primeiro fracaso. Do mesmo xeito que na distribución binomial, a probabilidade de éxito p, é a mesma para todos os ensaios.

Se X é unha variable aleatoria que segue unha distribución xeométrica de parámetro p, $X \sim Xeom(p)$, X é o número de éxitos ata o primeiro fracaso e os seus valores poden ser 0,1,2,...

Se consideramos o experimento aleatorio de lanzar unha moeda, a variable aleatoria que describe o número de caras antes de que saia a primeira cruz segue unha distribución xeométrica de parámetro $p = \frac{1}{2}$.

A función de masa de probabilidade de X describe a probabilidade de obter i éxitos ata o primeiro fracaso e vén dada por

$$P\{X = i\} = (1-p)p^i, \quad i = 0, 1, 2, ...$$

Supoñamos que $X \sim Xeom(p)$ e calculemos a súa media.

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} iP\{X = i\} = \sum_{i=1}^{\infty} i(1-p)p^i = (1-p)\sum_{i=1}^{\infty} ip^i = (1-p)p\sum_{i=1}^{\infty} ip^{i-1} \\ &= (1-p)p\frac{1}{(1-p)^2} = \frac{p}{1-p}, \end{aligned}$$

xa que $\sum_{i=1}^{\infty} ip^{i-1} = \frac{d}{dp}\left(\sum_{i=1}^{\infty} p^i\right)$ e $\sum_{i=1}^{\infty} p^i = \frac{p}{1-p} < \infty,\ p < 1$. Obtemos

$$E[X] = \frac{p}{1-p}. \tag{2.3}$$

A **distribución de Poisson** é unha distribución discreta que describe o número de veces que ocorre un suceso nun intervalo, ben pode ser de tempo ou de espazo. O parámetro λ, asociado á distribución, representa o número medio de sucesos por unidade de tempo ou espazo e ademais, λ é positivo, constante e non depende do intervalo de tempo ou espazo.

Se X é unha variable aleatoria que segue unha distribución de Poisson de parámetro λ, $X \sim Pois(\lambda)$, X é o número de sucesos que ocorre nun intervalo e pode tomar os valores $0, 1, 2, \ldots$.

Por exemplo, nunha sala de urxencias a variable aleatoria que describe o número de pacientes que son atendidos cada hora, a unha taxa media de 5 pacientes por hora, segue unha distribución de Poisson de parámetro $\lambda = 5$.

A función de masa de probabilidade de X indica a probabilidade de que un suceso ocorra i veces nun intervalo continuo de tempo ou espazo. A súa expresión vén dada por

$$P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, 2, ...$$

Supoñamos que $X \sim Pois(\lambda)$ e calculemos a súa media.

$$E[X] = \sum_{i=0}^{\infty} iP\{X = i\} = \sum_{i=1}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = \lambda e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda,$$

onde na cuarta igualdade reindexamos tomando $k = i - 1$. Obtemos

$$E[X] = \lambda. \tag{2.4}$$

Un resultado que empregaremos en capítulos seguintes é a aproximación da distribución binomial pola distribución de Poisson, sempre e cando os parámetros da binomial, n e p, cumpran certos requisitos que veremos a continuación.

Proposición 2.26. *Se X é unha variable aleatoria que segue unha distribución binomial de parámetros n e p, con $\lambda = p \cdot n$, entón, cando n tende a infinito, a probabilidade de que X tome un valor i, é dicir, $P(X = i)$, achégase a*

$$e^{-\lambda}\frac{\lambda^i}{i!} \quad \textit{para } i = 0, 1, ...$$

Este valor coincide coa probabilidade de que unha variable que segue unha distribución de Poisson de parámetro λ tome o valor i.

Demostración. Tomemos X tal que $X \sim Bin(n, p)$, con $p = \frac{\lambda}{n}$ e $\lambda > 0$. Entón, para cada $i \in \{1, 2, ..., n\}$

$$\begin{aligned} P\{X = i\} &= \binom{n}{i}p^i(1-p)^{n-i} = \frac{n(n-1)\cdot\cdots\cdot(n-i+1)(n-i)(n-i-1)\cdot\cdots\cdot 1}{i!(n-i)(n-i-1)\cdot\cdots\cdot 1}\cdot p^i(1-p)^{n-i} \\ &= \frac{n(n-1)\cdot\cdots\cdot(n-i+1)}{i!}\cdot\frac{\lambda^i}{n^i}(1-\tfrac{\lambda}{n})^{n-i} = \frac{\lambda^i}{i!}(1-\tfrac{1}{n})\cdot\cdots\cdot(1-\tfrac{i-1}{n})(1-\tfrac{\lambda}{n})^n(1-\tfrac{\lambda}{n})^{-i}, \end{aligned}$$

onde na segunda igualdade empregamos a definición do número combinatorio.

Agora, calculando os seguintes límites cando $n \to \infty$

- Para cada $k = 0, ..., i-1$, temos $\lim_{n\to\infty}\left(1 - \frac{k}{n}\right) = 1$.
- $\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$.
- $\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-i} = 1$.

Polo tanto, facendo tender $n \to \infty$, obtense

$$P(X = i) \to \frac{\lambda^i}{i!} \cdot e^{-\lambda},$$

chegando ao resultado que queríamos demostrar. □

Notemos que p ten que ser "pequeno" pois n é "grande" e $np = \lambda$ debe de permanecer constante.

Dicimos que X é unha variable aleatoria que segue unha **distribución exponencial** se a súa función de densidade f é

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{se } x \geq 0, \\ 0, & \text{noutro caso.} \end{cases}$$

A partir da función de densidade podemos obter a función de distribución de X que vén dada por

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{se } x \geq 0, \\ 0, & \text{noutro caso.} \end{cases}$$

Supoñamos que $X \sim Exp(\lambda)$ e calculemos a súa media.

$$E[X] = \int_0^\infty x f(x)\, dx = \int_0^\infty x\lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty x e^{-\lambda x}\, dx = \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda}.$$

Obtemos

$$E[X] = \frac{1}{\lambda}. \tag{2.5}$$

A distribución de exponencial é unha distribución continua que pode describir o tempo que pasa entre dous sucesos consecutivos.

A distribución exponencial posúe o que se coñece como "falta de memoria", recollida no seguinte resultado.

Proposición 2.27. *Sexa $X \sim Exp(\lambda)$ entón verifícase que*

$$P\{X > n + m \mid X > n\} = P\{X > m\}, \quad \forall\ n, m \geq 0.$$

Demostración.

$$\begin{aligned} P\{X > n + m \mid X > n\} &= \frac{P\{X > n + m, X > n\}}{P\{X > n\}} = \frac{P\{X > n + m\}}{P\{X > n\}} = \frac{1 - F(n + m)}{1 - F(n)} \\ &= \frac{e^{-(n+m)\lambda}}{e^{-n\lambda}} = e^{-m\lambda} = P\{X > m\}. \end{aligned}$$

□

Seguindo co exemplo anterior, a variable aleatoria que describe o tempo de espera dun paciente que acaba de chegar á sala de urxencias segue unha distribución exponencial de media $\frac{1}{5}$, é dicir 12 minutos.

A "falta de memoria" vén a dicir, que da igual o tempo que leve esperando o paciente a ser atendido. Se levase 10 minutos esperando a probabilidade de que tarde 5 minutos máis é a mesma que se acabase de chegar.

Se X é unha variable aleatoria que segue unha **distribución uniforme continua** no intervalo $[a, b]$, $X \sim U(a, b)$, X toma valores no intervalo $[a, b]$ e a súa función de densidade é constante en todo o intervalo.

A función de densidade de X describe a densidade de probabilidade, é dicir, como se distribúe a probabilidade ao longo de cada intervalo e vén dada por

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{se } a \leq x \leq b, \\ 0, & \text{noutro caso.} \end{cases}$$

A partir da función de densidade obtemos a función de distribución, que describe a probabilidade de que a variable aleatoria tome un valor menor ou igual que x.

$$F(x) = \begin{cases} 0, & \text{se } x < a, \\ \frac{x-a}{b-a}, & \text{se } a \leq x \leq b, \\ 0, & \text{se } x > b. \end{cases}$$

Notemos que a probabilidade de cada intervalo non depende de onde se localice o intervalo dentro de $[a, b]$, se non que depende da lonxitude do intervalo que consideremos.

En efecto, se consideramos un intervalo $[c, d] \subset [a, b]$, logo a probabilidade de que a variable aleatoria tome un valor no intervalo $[c, d]$

$$P\{c \leq X \leq d\} = \int_c^d f(x)\, dx = \int_c^d \frac{1}{b-a}\, dx = \frac{d-c}{b-a}.$$

Supoñamos que $X \sim U(a, b)$ e calculemos a súa media.

$$E[X] = \int_a^b x f(x)\ dx = \int_a^b \frac{x}{b-a}\ dx = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2},$$

Obtemos

$$E[X] = \frac{a+b}{2}. \tag{2.6}$$

Imaxinemos que pechamos os ollos e pousamos o noso dedo sobre un punto dunha regra de 30 centímetros. A variable aleatoria X que describe a posición do punto sinalado segue unha

distribución uniforme no intervalo $[0, 30]$. Ademais, a probabilidade de que o punto caia nun determinado intervalo de $[0, 30]$, depende só da lonxitude deste. Canto máis pequeno sexa o intervalo, menos probabilidade de que o punto caia nel.

Con este capítulo presentamos os conceptos teóricos sobre a probabilidade e as variables aleatorias que se empregarán nos seguintes capítulos.

Capítulo 3

Procesos estocásticos

3.1. Introdución

Moitos mecanismos aleatorios interveñen na bioinformática, en concreto, no estudo de secuencias de ADN. Neste capítulo introducimos os procesos estocásticos facendo fincapé nos procesos de Poisson. A información recollida neste capítulo ten como referencia o libro [6].

Definición 3.1. Sexa $(\Omega, \mathcal{F}, P)$ un espazo de probabilidade. Un **proceso estocástico** é unha colección $\{X_t,\ t \in T\}$ de variables aleatorias, onde para cada t do **conxunto de índices** T, X_t definida como

$$\begin{aligned} X_t \ &: \ \Omega \longrightarrow \mathbb{R} \\ &\quad\ \omega \longmapsto X_t(\omega) \end{aligned}$$

é unha variable aleatoria e para cada $\omega \in \Omega$ $X_t(\omega)$ é un **estado do proceso no instante** t. O conxunto de estados do proceso chámase **espazo de estados**.

Estamos considerando un proceso, unha secuencia de eventos, que evoluciona de forma aleatoria en cada instante. No obstante, aínda que o parámetro t soe ser o tempo, tamén pode representar calquera outro parámetro de evolución, como a distancia nunha secuencia de ADN.

Distinguimos diferentes tipos de procesos estocásticos en función do seu espazo de estados. Nun proceso estocástico, se as variables aleatorias toman un número finito ou infinito numerable de valores dicimos que ten **espazo de estados discreto**, mentres que se as variables aleatorias toman valores en $\mathbb{R}$ dicimos que ten **espazo de estados continuo**.

Se o conxunto de índices T é un conxunto numerable, diremos que $\{X_t,\ t \in T\}$ é un **proceso estocástico en tempo discreto**, mentres que se $T = \mathbb{R}$, dicimos que $\{X_t,\ t \in T\}$ é un **proceso estocástico en tempo continuo**.

Por exemplo, o proceso que describe o número de clientes que hai nun banco en cada instante entre as 10 e as 12 da mañá é un proceso estocástico en tempo continuo con espazo de estados discreto. Para cada instante de tempo comprendido entre as 10 e as 12 horas temos unha variable aleatoria que describe o número de clientes que hai no banco nese instante, e polo tanto, toma un número finito de valores.

Un proceso estocástico en tempo continuo $\{X_t,\ t \in T\}$ dise que ten **incrementos independentes** se para todo $t_0 < t_1 < t_2 < ... < t_n$, as variables aleatorias

$$X_{t_1} - X_{t_0},\ X_{t_2} - X_{t_1}, ...,\ X_{t_n} - X_{t_{n-1}},$$

son independentes. Isto é, dicimos que un proceso estocástico ten incrementos independentes se os cambios no proceso durante intervalos de tempo disxuntos son independentes.

Un proceso estocástico $\{X_t,\ t \in T\}$ ten **incrementos estacionarios** se $X_{t+s} - X_t$ ten a mesma distribución para todo t. Isto significa que, a distribución das variables depende só da lonxitude do intervalo de tempo.

A **propiedade markoviana** di que a distribución condicional dun estado futuro no tempo $t+s$, dado o estado presente no tempo s e todos os estados pasados, depende só do estado presente e é independente dos estados pasados, é dicir,

$$P\{X_{t+s} = j \mid X_s = i, X_u = x_u, 0 \leq u < s\} = P\{X_{t+s} = j \mid X_s = i\}.$$

Coñécese como a "falta de memoria".

Un tipo de proceso estocástico en tempo continuo e espazo de estados discreto son os procesos de conteo que permiten modelizar situacións nas que queremos contar a aparición de eventos ao longo do tempo.

Definición 3.2. Un proceso estocástico $\{X_t,\ t \geq T\}$ dise **proceso de conteo** se X_t representa o número total de eventos que ocorreron ata o instante t, satisfacendo:

(i) $X_t \geq 0$.
(ii) X_t é un valor enteiro.
(iii) Se $s < t$, logo $X_s \leq X_t$.
(iv) Para $s < t$, $X_t - X_s$ correspóndese co número de eventos que ocorreron no intervalo $(s, t]$.

Dicimos que un proceso de conteo ten **incrementos independentes** se as variables aleatorias que describen o número de eventos que ocorren en intervalos disxuntos son independentes. É dicir, se para todo $t_0 < t_1 < t_2 < ... < t_n$, as variables aleatorias

$$X_{t_1} - X_{t_0},\ X_{t_2} - X_{t_1}, ...,\ X_{t_n} - X_{t_{n-1}},$$

son independentes.

Por outro lado, dicimos que un proceso de conteo posúe **incrementos estacionarios** se a distribución do número de eventos que ocorren en calquera intervalo de tempo depende só da lonxitude de dito intervalo. É dicir, se $X_{t+s} - X_t$ ten a mesma distribución para todo t.

3.2. Procesos de Poisson

Un tipo de proceso de conteo son os procesos de Poisson. Estes procesos caracterízanse porque os eventos aparecen de forma independente e ocorren a unha taxa constante no tempo. Ademais, os procesos de Poisson, verifican a propiedade markoviana. A continuación presentamos a súa definición.

Definición 3.3. Un proceso de conteo $\{X_t,\ t \geq 0\}$ dise **proceso de Poisson** con taxa $\lambda,\ \lambda > 0$, se verifica

(i) $X_0 = 0$.
(ii) Posúe incrementos independentes.
(iii) O número de eventos en calquera intervalo de lonxitude t segue unha distribución de Poisson de media λt, isto é, para todo $s, t \geq 0$

$$P\{X_{t+s} - X_s = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, ...$$

Pola última condición, como $X_t \sim \text{Pois}(\lambda t)$, temos que

$$E[X_t] = \lambda t.$$

Polo tanto, o parámetro λ é o número medio de eventos que ocorren por unidade de tempo nun proceso de Poisson.

Teorema 3.4. *O proceso de conteo $\{X_t,\ t \geq 0\}$ é un proceso de Poisson con taxa λ, $\lambda > 0$, se, e só se verifica*

(i) $X_0 = 0$.
(ii) Posúe incrementos independentes e estacionarios.
(iii) $P\{X_h = 1\} = \lambda h + o(h)$ *e* $P\{X_h \geq 2\} = o(h)$, *onde* $\lim_{h\to 0} \frac{o(h)}{h} = 0$.

Demostración. Comecemos probando que baixo as condicións dadas no Teorema 3.4 o proceso de conteo verifica a definición de proceso de Poisson. Notemos que só precisamos verificar a condición (iii), de que o número de eventos en calquera instante de lonxitude t segue unha distribución de Poisson de media λt, xa que o resto verifícanse por hipótese.

Empregaremos a notación

$$P_n(t) = P\{X_t = n\}, \quad \forall n \in \mathbb{N},$$

e demostraremos que $X_t \sim Pois(\lambda t)$, é dicir, que

$$P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad \forall n \in \mathbb{N},$$

por indución en n.

Comecemos probando o resultado para $n = 0$.

$$\begin{aligned} P_0(t+h) &= P\{X_{t+h} = 0\} \\ &= P\{X_t = 0,\ X_{t+h} - X_t = 0\} \\ &= P\{X_t = 0\}\, P\{X_{t+h} - X_t = 0\} \\ &= P_0(t)[1 - \lambda h + o(h)] \\ &= P_0(t) - \lambda h P_0(t) + o(h), \end{aligned}$$

onde a terceira igualdade vén dos incrementos independentes, e a cuarta igualdade, pola estacionalidade de eventos e que $P\{X_h = 0\} = 1 - \lambda h + o(h)$, sabendo que

$$P\{X_h \geq 2\} = 1 - P\{X_h = 0\} - P\{X_h = 1\},$$

e ademais,

$$P\{X_h = 1\} = \lambda h + o(h) \quad \text{e} \quad P\{X_h \geq 2\} = o(h).$$

Así, reordenando termos e dividindo entre h obtemos

$$\frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t) + \frac{o(h)}{h},$$

e tomando límites cando $h \to 0$ temos

$$P_0'(t) = -\lambda P_0(t).$$

Logo, integrando a expresión

$$\int \frac{P_0'(t)}{P_0(t)}\, dt = \int -\lambda\, dt$$

obtemos, para unha constate c_1,

$$\log P_0(t) = -\lambda t + c_1,$$

equivalentemente, para unha constante c_2,

$$P_0(t) = c_2 e^{-\lambda t}.$$

Como $X_0 = 0$, logo $P_0(0) = 1$. Impoñendo esta condición inicial á ecuación anterior chegamos ao resultado

$$P_0(t) = e^{-\lambda t}, \tag{3.1}$$

quedando probado para $n = 0$.

Probémolo para $n = 1$. Tomamos $n \geq 1$ e procedendo de forma similar,

$$\begin{aligned} P_n(t+h) &= P\{X_{t+h} = n\} \\ &= P\{X_t = n,\ X_{t+h} - X_t = 0\} \\ &\quad + P\{X_t = n-1,\ X_{t+h} - X_t = 1\} \\ &\quad + P\{X_{t+h} = n,\ X_{t+h} - X_t \geq 2\}. \end{aligned}$$

Como $P\{X_h \geq 2\} = o(h)$, o último termo da expresión anterior é un $o(h)$. Logo, polos incrementos estacionarios e independentes obtemos

$$\begin{aligned} P_n(t+h) &= P_n(t)P_0(h) + P_{n-1}(t)P_1(h) + o(h) \\ &= P_n(t)(1-\lambda h) + P_{n-1}(t)\lambda h + o(h). \end{aligned}$$

Reordenando termos e dividindo entre h, obtemos

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h},$$

e tomando límites cando $h \to 0$ temos

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t).$$

Reordenando e multiplicando por $e^{\lambda t}$

$$e^{\lambda t}[P_n'(t) + \lambda P_n(t)] = \lambda e^{\lambda t} P_{n-1}(t),$$

é dicir,

$$\frac{d}{dt}(e^{\lambda t} P_n(t)) = \lambda e^{\lambda t} P_{n-1}(t). \tag{3.2}$$

Aplicando a ecuación (3.2) para o caso $n = 1$ e usando o resultado dado en (3.1) temos

$$\frac{d}{dt}(e^{\lambda t} P_1(t)) = \lambda,$$

integrando e despexando $P_1(t)$ obtemos, para unha constante k_1,

$$P_1(t) = (\lambda t + k_1)e^{-\lambda t}.$$

Como $X_0 = 0$, temos que $P_1(0) = 0$. Impoñendo esta condición inicial á ecuación anterior chegamos ao resultado

$$P_1(t) = \lambda t e^{-\lambda t},$$

quedando probado para $n = 1$.

Supoñamos agora que o resultado é certo para $n-1$, é dicir, que

$$P_{n-1}(t) = e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}.$$

Logo, por (3.2), aplicando a hipótese de indución temos

$$\frac{d}{dt}(e^{\lambda t}P_n(t)) = \frac{\lambda(\lambda t)^{n-1}}{(n-1)!},$$

integrando obtemos, para un k_2 constante,

$$e^{\lambda t}P_n(t) = \frac{(\lambda t)^n}{n!} + k_2.$$

Como $X_0 = 0$, temos que $P_n(0) = 0$ para todo $n \geq 1$. Impoñendo esta condición inicial á ecuación anterior chegamos ao resultado que queríamos demostrar

$$P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

Reciprocamente, probemos que un proceso de Poisson verifica as condicións do Teorema 3.4. Como consecuencia da distribución de Poisson do número de eventos en cada instante, (iii), os procesos de Poisson tamén posúen incrementos estacionarios. Logo, só queda demostrar que

$$P\{X_h = 1\} = \lambda h + o(h) \quad \text{e} \quad P\{X_h \geq 2\} = o(h).$$

Como $X_h \sim \text{Pois}(\lambda h)$, por un lado,

$$P\{X_h = 1\} = e^{-\lambda h}\lambda h = \lambda h(1 - \lambda h + o(h)) = \lambda h - \lambda^2 h^2 + o(h) = \lambda h + o(h),$$

aplicando na segunda igualdade que $e^{-\lambda h} = 1 - \lambda h + o(h)$, xa que

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots$$

Por outro lado,

$$\begin{aligned} P\{X_h \geq 2\} &= 1 - P\{X_h = 0\} - P\{X_h = 1\} \\ &= 1 - e^{-\lambda h} - e^{-\lambda h}\lambda h \\ &= 1 - (1 - \lambda h + o(h)) - \lambda h(1 - \lambda h + o(h)) \\ &= \lambda h + o(h) - \lambda h + \lambda^2 h^2 + o(h)\lambda h \\ &= o(h), \end{aligned}$$

aplicando que $P\{X_h = 0\} = e^{-\lambda h}$. □

Consideremos un proceso de Poisson e definimos a variable aleatoria T_n como o tempo entre o evento n-1-ésimo eo evento n-ésimo, para $n \geq 1$. A secuencia de variables aleatorias $\{T_n,\ n \geq 1\}$ é a **secuencia de tempos entre eventos** dun proceso de Poisson.

Proposición 3.5. *As variables aleatorias que describen o tempo entre eventos dun proceso de Poisson, $T_1, T_2, \dots$, son independentes e seguen unha distribución exponencial con parámetro λ.*

Demostración. Comecemos calculando a distribución de T_1. Notemos que $T_1 > t$ se, e só se, non ocorre ningún evento do proceso de Poisson no intervalo $[0, t]$, logo

$$P\{T_1 > t\} = P\{0 \text{ eventos en } [0,t]\} = P\{X_t = 0\} = e^{-\lambda t},$$

aplicando que $X_t \sim Pois(\lambda t)$. Así obtemos que T_1 segue unha distribución exponencial de parámetro λ.

Calculemos agora a distribución de T_2 (o tempo entre o primeiro e o segundo evento) condicionada a T_1

$$\begin{aligned} P\{T_2 > t \mid T_1 = s\} = P\{0 \text{ eventos en } (s, s+t] \mid 1 \text{ evento en } [0,s]\} &= P\{0 \text{ eventos en } (s, s+t]\} \\ = P\{X_{t+s} - X_s = 0\} = P\{X_t = 0\} &= e^{-\lambda t}, \end{aligned}$$

onde na segunda igualdade aplícase os incrementos independentes. Como consecuencia $P\{T_2 > t \mid T_1 = s\} = P\{T_2 > t\}$, así T_2 é independente de T_1. Ademais, como os incrementos son estacionarios, concluímos que T_2 tamén segue unha distribución exponencial de parámetro λ.

Repetindo o mesmo argumento para as seguintes variables, obtemos que $T_1, T_2, \dots$ son independentes e seguen unha distribución exponencial de parámetro λ, tal e como queríamos demostrar. □

A suposición de incrementos estacionarios e independentes implica que en calquera instante o proceso "volve a comezar" probabilisticamente, é dicir, o proceso non ten memoria.

Como vimos no capítulo 2 a distribución exponencial caracterízase pola falta de memoria, polo tanto, é de esperar que os tempos entre eventos sexan independentes e sigan unha distribución exponencial, tal e como se demostrou.

Con este capítulo rematamos co marco teórico necesario para resolver os problemas tratados ao longo dos seguintes capítulos.

Capítulo 4

Ensamblado de secuencias

4.1. Introdución

Para analizar o material xenético é necesario obter o ADN que se quere estudar. A **secuenciación** do ADN é unha técnica que permite identificar e ordenar os nucleótidos que forman parte do ADN orixinal, para representalos en forma dunha ou varias secuencias.

O **xenoma** é o conxunto de todo o material xenético dun organismo. O noso material xenético está contido en 23 pares de longas moléculas de ADN chamadas **cromosomas**. Ademais, o xenoma humano consta de aproximadamente tres mil millóns de nucleótidos, polo que esta extensa lonxitude supón unha barreira á hora de secuenciar o xenoma por completo.

A secuenciación *shotgun* é unha técnica empregada na secuenciación do xenoma, que consiste en romper o ADN en pequenos fragmentos que son secuenciados posteriormente. Acto seguido, os fragmentos son ensamblados (unidos) computacionalmente para reconstruír o xenoma de maneira completa, tal e como podemos ver na Figura 4.1.

Para a secuenciación dos pequenos fragmentos empréganse técnicas de secuenciación, que permiten ler unha serie de nucleótidos consecutivos dende un dos extremos de cada fragmento. As secuencias que se obteñen como resultado denomínanse **lecturas**, representadas na Figura 4.2 en cor azul. A lonxitude das lecturas depende da técnica de secuenciación empregada.

Normalmente, descoñécese a localización e orientación exactas de cada lectura dentro do xenoma orixinal. Sen embargo, ao repetir os procesos de fragmentación e secuenciación as lecturas presentan solapamentos entre elas, tal e como podemos ver na Figura 4.2, que permiten que sexan aliñadas e ensambladas empregando un software de ensamblado de secuencias.

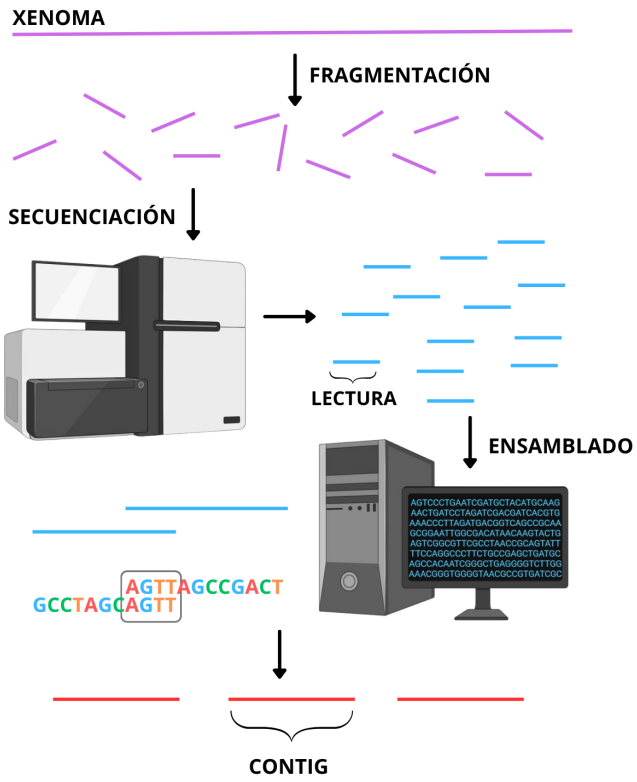

Figura 4.1: Representación da secuenciación shotgun creada empregando BioRender.com.

O ensamblado das lecturas solapadas dá lugar a secuencias de ADN máis longas chamadas **contigs**, representadas na Figura 4.2 en cor vermella, que agrupadas forman unha reconstrución do xenoma orixinal.

Precísanse moitas lecturas superpostas para construír con precisión cada secuencia. A **cobertura** é o número medio de veces que se secuencia cada nucleótido da secuencia orixinal, polo que, aparentemente coberturas altas deberían dar mellores resultados.

Definimos os seguintes parámetros:

- L_G = lonxitude total do xenoma.
- L = lonxitude de cada lectura.
- N = número total de lecturas obtidas.
- a = cobertura (número medio de veces que se secuencia cada nucleótido da secuencia orixinal)

Notemos que $N \cdot L$ representa a lonxitude total das lecturas. Polo tanto, a partir destes paráme-

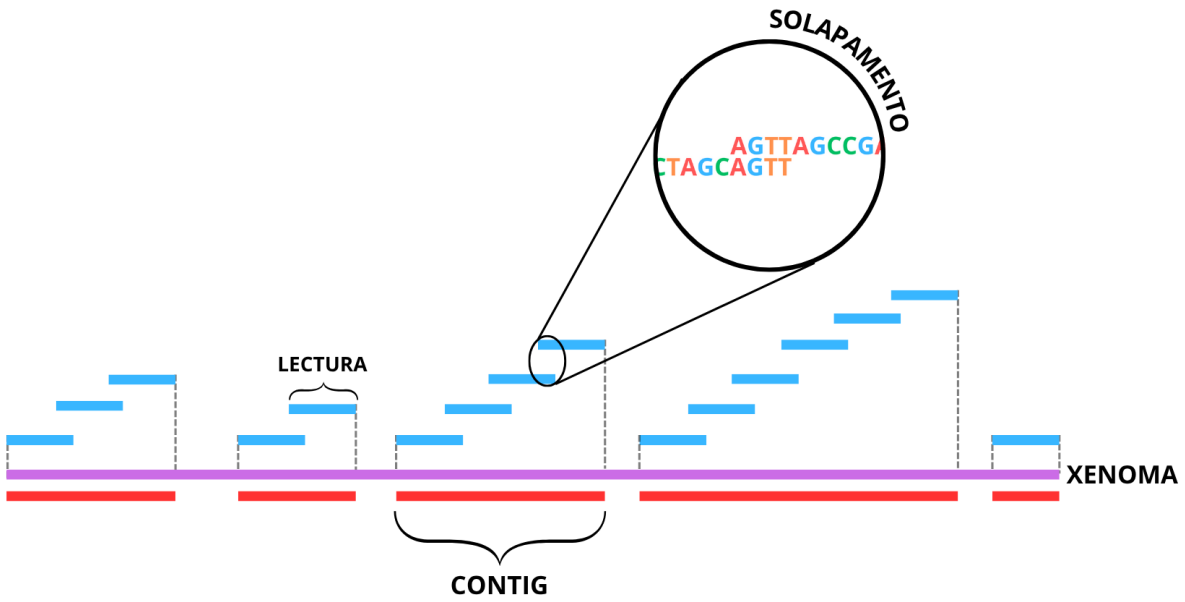

Figura 4.2: Representación da estrutura dos contigs.

tros, temos a seguinte expresión para a cobertura

$$a = \frac{N \cdot L}{L_G}.$$

Adoitase expresar a cobertura como "aX", o que significa que, se a lonxitude orixinal do xenoma é L_G entón a lonxitude total de lecturas obtidas é $a \cdot L_G$.

Polo xeral, a maior cobertura obtemos un menor número de contigs, xa que se formarán un maior número de solapamentos, e ademais, os contigs serán máis longos, xa que estarán compostos por un maior número de lecturas.

Este procedemento, que permite reconstruír longas secuencias de ADN, baséase no estudo das probabilidades e nos procesos estocásticos, que permiten analizar os resultados obtidos e estudar como mellorar a similitude co xenoma orixinal. Preséntanse cuestións como a proporción esperada do xenoma cuberto, o número esperado de contigs ou a lonxitude esperada dos contigs, problemas que trataremos nas seguintes seccións.

Ao longo deste capítulo traballarase cun modelo probabilista máis simple que a realidade biolóxica, asumindo certas propiedades para facilitar a súa análise. Isto permite identificar tamén os posibles problemas ou inexactitudes que pode presentar. Esta análise dá paso ao refinamento do

modelo, dando lugar a modelos máis complexos e realistas.

Tanto nesta introdución como nas seccións que seguen, o material empregado ten como referencia o libro [3].

4.2. Proporción esperada do xenoma cuberto por contigs

Nesta sección búscase determinar a proporción esperada do xenoma cuberto por contigs. Isto tradúcese en calcular a probabilidade de que un punto calquera P do xenoma estea cuberto polo menos por unha lectura. É dicir, a probabilidade de que polo menos unha lectura teña o seu extremo esquerdo dento dun intervalo de lonxitude L á esquerda do punto P como podemos ver na Figura 4.3.

Polo tanto traballaremos co extremo esquerdo de cada lectura, é dicir, co comezo de cada lectura, que consideraremos como un punto. Logo debemos de calcular a probabilidade de que polo menos un extremo esquerdo caia no intervalo $(P - L, L)$.

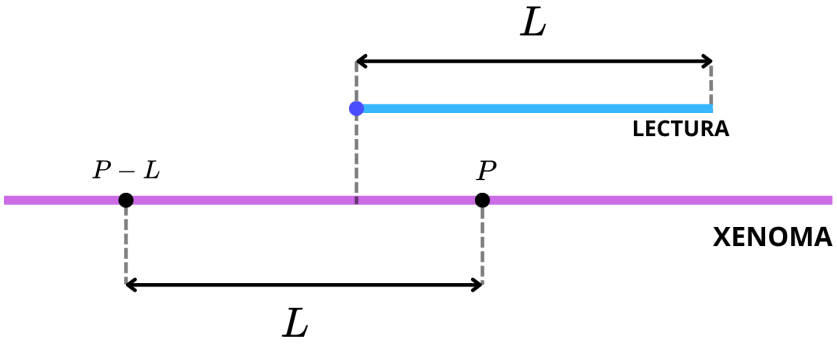

Figura 4.3: Punto P cuberto por unha lectura.

Realizaremos este cálculo en función da cobertura a, considerando un xenoma de lonxitude total L_G e un total de N lecturas, cada unha de lonxitude L con $L < L_G$. Antes de proceder é preciso facer unhas consideracións iniciais.

En primeiro lugar, vexamos que posicións poden ocupar estes puntos. Nótese que, dentro dun cromosoma unha lectura pode comezar en calquera posición excepto nas últimas $L-1$ posicións do cromosoma pois "sairíase fóra". Polo tanto, se o xenoma ten c cromosomas as posicións excluídas aumentan a $c(L-1)$.

Cando o numero posicións rechazadas é desprezable fronte á lonxitude total do xenoma L_G, dicimos que ignoramos os efectos finais (como é o caso do xenoma humano) e consideramos que cada extremo esquerdo pode ocupar calquera posición do intervalo $[0, L_G]$.

Se supoñemos que cada lectura se escolle ao azar dentro do xenoma e que son independentes entre elas, entón, os extremos esquerdos de cada lectura seguen unha distribución uniforme no intervalo $[0, L_G]$.

Polo tanto, se $f : \mathbb{R} \to \mathbb{R}$ definida como $f(x) = \frac{1}{L_G}$, para cada $x \in [0, L_G]$ é a función de densidade da distribución uniforme en $[0, L_G]$, entón a probabilidade de que un extremo esquerdo caia no intervalo $(P - L, P)$ é

$$\int_{P-L}^{P} f(t)\, dt = \int_{P-L}^{P} \frac{1}{L_G}\, dt = \frac{L}{L_G}.$$

Notemos que, esta probabilidade é a mesma para todo intervalo de lonxitude L.

Consideramos como un "éxito" que un extremo esquerdo caia nun intervalo de lonxitude L e definimos a variable aleatoria X_L, que describe o número de extremos esquerdos que caen nun intervalo de lonxitude L de entre N lecturas. É dicir, X_L describe o número de éxitos en N ensaios con probabilidade de éxito $\frac{L}{L_G}$. Polo tanto, X_L segue unha distribución binomial con probabilidade de éxito $\frac{L}{L_G}$ e número de intentos N, $X_L \sim Bin\left(N, \frac{L}{L_G}\right)$.

Pola Proposición 2.26, a distribución binomial pode aproximarse a unha distribución de Poisson cando o número de intentos é "grande" e a probabilidade de éxito é "pequena".

No noso caso, para unha cobertura fixa a, se $N \to \infty$

$$\frac{L}{L_G} = \frac{\frac{a \cdot L_G}{N}}{L_G} = \frac{a}{N} \to 0,$$

aplicando na segunda igualdade que $a = \frac{N \cdot L}{L_G}$. Logo se N é "grande", $\frac{L}{L_G}$ é "pequena".

Entón, podemos considerar que a variable aleatoria X_L segue aproximadamente unha distribución de Poisson de media $N\frac{L}{L_G}$, $X \sim Pois(\frac{N \cdot L}{L_G})$.

Consideramos a colección de variables aleatorias $\{X_L, L \geq 0\}$, onde para cada parámetro L, que representa unha distancia na secuencia de ADN, a variable aleatoria X_L describe o número de extremos ata a distancia L. Notemos que se trata dun proceso de Poisson.

En primeiro lugar, trátase dun proceso de tempo continuo e espazo de estados discreto. Ademais, sabemos que o número medio de eventos por unidade de distancia, é dicir, o número medio de extremos que caen por unidade de distancia é $\frac{N}{L_G}$. Polo tanto, trátase dun proceso de conteo onde os eventos ocorren a unha taxa constate $\lambda = \frac{N}{L_G} > 0$.

Vexamos que se verifican as condicións da definición:

(i) Claramente o proceso empeza en cero, xa que para unha distancia nula non temos nada, logo non temos ningún extremo esquerdo, é dicir, $X_0 = 0$.

(ii) Por hipótese os extremos esquerdos son independentes, logo o número de extremos que caen en intervalos disxuntos son independentes. Dicimos que o proceso ten incrementos independentes.

(iii) Como xa vimos, a variable aleatoria X_L segue unha distribución de Poisson de media $\frac{N \cdot L}{L_G}$, é dicir, o número de extremos esquerdos que caen en calquera intervalo de lonxitude L segue unha distribución de Poisson de media λL onde λ é a taxa $\frac{N}{L_G}$. Polo tanto

$$P\{X_L = i\} = e^{-\lambda L} \frac{(\lambda L)^i}{i!}, \quad i = 0, 1, ...$$

Tal e como esperabamos, o proceso cumpre a definición de proceso de Poisson. É un proceso de conteo $\{X_L, L \geq 0\}$ con taxa constante $\lambda = \frac{N}{L_G} > 0$, empeza en cero, posúe incrementos independentes e o número de eventos en calquera intervalo de lonxitude L segue unha distribución de Poisson de media λL.

Ademais, como os extremos esquerdos seguen unha distribución uniforme no intervalo $[0, L_G]$, a distribución das variables só depende do intervalo considerado, é dicir, posúe incrementos estacionarios.

Polo tanto, a probabilidade de que un punto calquera P estea cuberto polo menos por unha lectura é a mesma que a probabilidade de que polo menos unha lectura teña o seu extremo esquerdo dentro dun intervalo de lonxitude L á esquerda deste punto P, é dicir

$$P\{X_P - X_{P-L} \geq 1\} = P\{X_L \geq 1\} = 1 - P\{X_L = 0\} = 1 - e^{-\lambda L} \frac{(\lambda L)^0}{0!} = 1 - e^{-\lambda L},$$

aplicando na primeira igualdade os incrementos estacionarios e na terceira que $X \sim Poiss(\lambda L)$.

Finalmente, como $\lambda L = \frac{N \cdot L}{L_G} = a$, logo a proporción esperada do xenoma cuberto por contigs en función da cobertura a é

$$\textbf{Proporción esperada do xenoma cuberto por contigs} = 1 - e^{-a}.$$

4.3. Número esperado de contigs

Nesta sección buscamos calcular o número esperado de contigs. Notemos que nun contig podemos diferenciar unha lectura do resto de lecturas que o forman. Trátase da lectura final que denotaremos coa etiqueta F, para diferenciala do resto de lecturas, tal e como podemos ver na Figura 4.4.

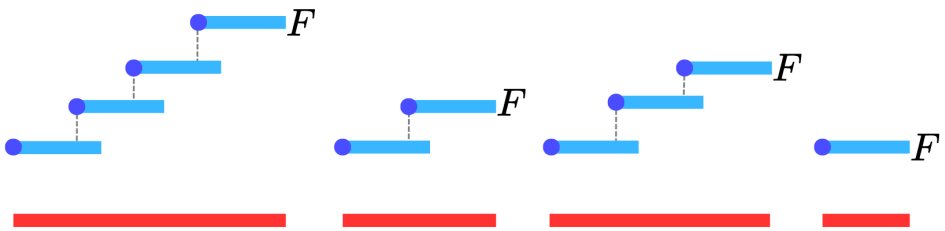

Figura 4.4: Representación das lecturas que forman os contig.

Como vemos, cada contig ten unha única lectura do tipo F. Polo tanto, como este tipo de lecturas é única en cada contig, o número de esperado de contigs será o mesmo que o número esperado de lecturas do tipo do F. Calculemos a probabilidade de que unha lectura sexa do tipo F.

Notemos que, a continuación dunha lectura tipo F, hai un "salto" ao seguinte contig pola falta de solapamento. Logo, unha lectura é do tipo F, é dicir, unha lectura é a lectura final dun contig se ningún extremo esquerdo de calquera outra lectura cae dentro da lectura F.

Como o espazo que ocupa unha lectura de tipo F é un intervalo de lonxitude L, a probabilidade buscada é a probabilidade de que o número de extremos esquerdos que caen nun intervalo de lonxitude L sexa 0, é dicir, que a variable aleatoria X_L, definida na sección anterior, tome o valor 0.

Como vimos que X_L segue unha distribución de Poisson de media λL onde $\lambda = \frac{N}{L_G}$, obtemos

$$P\{X_L = 0\} = e^{-\lambda L} = e^{-\lambda L}\frac{(\lambda L)^0}{0!} = e^{-\lambda L}.$$

Aplicando que $\lambda L = \lambda = \frac{N}{L_G}L = a$, a probabilidade de que unha lectura sexa do tipo F é e^{-a}.

Consideramos unha lectura do tipo F como un "éxito" e definimos a variable aleatoria Y que describe o número de lecturas tipo F entre un total de N lecturas, é dicir, Y describe o número de éxitos en N ensaios. Polo tanto, a variable aleatoria Y segue unha distribución binomial de N intentos e probabilidade de éxito e^{-a} , $Y \sim Bin(N, e^{-a})$.

Finalmente, o número esperado de contigs é a esperanza da variable aleatoria Y, que como segue unha distribución binomial, pola ecuación (2.2) obtemos

$$E[Y] = Ne^{-a}.$$

$$\textbf{Número esperado de contigs} = Ne^{-a}.$$

4.4. Tamaño esperado dos contigs

Un contig está formado por múltiples lecturas solapadas polo que o seu tamaño dependerá do número de lecturas que o formen e como estean colocadas. Recordemos que consideramos que todas as lecturas teñen a mesma lonxitude L.

Notemos que, se a distancia entre o extremo esquerdo dunha lectura e o extremo esquerdo da seguinte lectura é menor que L, significa que, a seguinte lectura comeza antes de que a primeira lectura remate. Entón, hai solapamento. Pola contra, se a distancia entre o extremo esquerdo dunha lectura e o extremo esquerdo da seguinte lectura é maior que L, non hai solapamento e remata o contig, tan el como podemos ver na Figura 4.5.

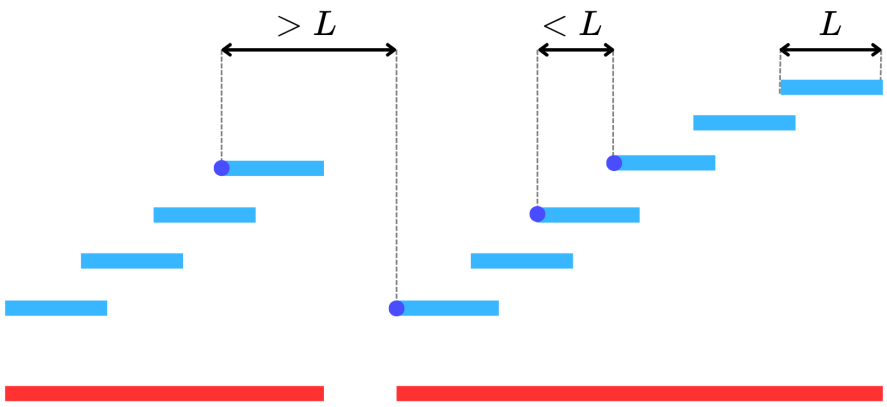

Figura 4.5: Representación das distancias entre lecturas.

Deste xeito, a lonxitude total esperada dun contig sería a lonxitude da última lectura, máis a suma das distancias esperadas entre os extremos esquerdos consecutivos do resto das lecturas solapadas.

A lonxitude da última lectura é L (igual que o resto de lecturas), mais, tanto o número de solapamentos como a distancia á que se producen, son datos descoñecidos, polo tanto é preciso calculalos.

Recordemos que na sección 4.2 obtivemos o proceso de Poisson $\{X_L, L \geq 0\}$ onde para cada parámetro L, que representa unha distancia na secuencia de ADN, a variable aleatoria X_L describe o número de extremos ata a distancia L.

Consideremos agora a variable aleatoria T_n para algún $n \geq 1$ que describe a distancia entre o $n-1$-ésimo evento e o n-ésimo evento do proceso de Poisson $\{X_L, L \geq 0\}$. É dicir, cada variable aleatoria T_n, $n \geq 1$, describe a distancia entre extremos esquerdos consecutivos.

Como xa sabemos, prodúcese un solapamento cando a distancia entre dous extremos esquerdos consecutivos é menor que a lonxitude destas, é dicir, cando a distancia entre dous eventos consecutivos do proceso de Poisson é menor que L.

Pola Proposición 3.5 as variables aleatorias $T_1, T_2, ...,$ son independentes e seguen unha distribución exponencial de parámetro $\lambda = \frac{N}{L_G}$. Logo a probabilidade de solapamento, para calquera $n \geq 1$, é

$$P\{0 < T_n \leq L\} = \int_0^L \lambda e^{-\lambda x}\, dx = \left[1 - e^{-\lambda x}\right]_0^L = 1 - e^{-\lambda L}.$$

Como $\lambda = \frac{N \cdot L}{L_G} = a$, obtemos que a probabilidade de solapamento é $1 - e^{-a}$.

Consideramos un solapamento como un "éxito" e definimos a variable aleatoria Z que describe o número de solapamentos nun contig. Como a falta de solapamento, é dicir, o "fracaso", marca o fin dun contig, Z describe o número de éxitos antes do primeiro fracaso. Logo a variable aleatoria Z segue unha distribución xeométrica con probabilidade de éxito $1 - e^{-a}$, $Z \sim Xeom(1 - e^{-a})$.

Como a variable aleatoria Z segue unha distribución xeométrica, aplicando a ecuación (2.3) obtemos que o número esperado de solapamentos nun contig é

$$E[Z] = e^a - 1.$$

A continuación calcúlase a distancia esperada entre lecturas solapadas.

Como xa vimos, as variables aleatorias $\{T_n, n \geq 1\}$ describen as distancias entre eventos consecutivos do proceso de Poisson. É dicir, describen as distancias entre extremos esquerdos de lecturas consecutivas. Ademais, para que se produza solapamento no extremo n-ésimo debe cumprirse que $0 < T_n \leq L$. Polo tanto, calculemos a esperanza da distancia entre os extremos esquerdos condicionada a que dita distancia estea entre 0 e L.

Esta medida vén dada, para calquera $n \geq 1$, por

$$E[T_n \mid 0 < T_n \leq L] = \int_0^L z h(v)\, dv,$$

onde, aplicando a ecuación (2.1), a función h vén dada pola expresión

$$h(v) = \frac{f(v)}{\int_0^L f(u)du} = \frac{\lambda e^{-\lambda v}}{1 - e^{-\lambda L}}, \quad \text{para } 0 < v \leq L.$$

Sabendo que $T_n \sim Exp(\lambda),\ \forall n \geq 1$, a integral do denominador é

$$\int_0^L \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda L}.$$

Obtemos que

$$E[T_n \mid 0 < T_n \leq L] = \int_0^L vh(v)\,dv = \int_0^L v \cdot \frac{\lambda e^{-\lambda v}}{1-e^{-\lambda L}}\,dv = \frac{\lambda}{1-e^{-\lambda L}} \int_0^L ve^{-\lambda v}\,dv$$
$$= \frac{\lambda}{1-e^{-\lambda L}} \left(-\frac{Le^{-\lambda L}}{\lambda} + \frac{1-e^{-\lambda L}}{\lambda^2} \right) = \frac{1}{\lambda} - \frac{L}{e^{\lambda L}-1},$$

Como $\lambda L = \frac{N \cdot L}{L_G} = a$, obtemos que a distancia esperada entre lecturas solapadas é

$$E[T_n | 0 < T_n \leq L] = \frac{1}{\lambda} - \frac{L}{e^a - 1}.$$

Finalmente, polo punto (ii) da Proposición 2.16, $E[X_1 \cdot X_2 \cdot \cdots \cdot X_n] = E[X_1]E[X_2] \cdot \cdots \cdot E[X_n]$ se $X_1, X_2, \ldots, X_n$ son variables aleatorias independentes. Claramente o número de solapamentos é independente da distancia entre solapamentos, é dicir, Z e T_n son variables aleatorias independentes, para calquera $n \geq 1$. Polo tanto, o tamaño esperado dun contig é a lonxitude da súa última lectura máis número esperado de solapamentos multiplicado pola distancia esperada entre lecturas solapadas, é dicir

$$E[Z] \cdot E[T_n | 0 < T_n \leq L] + L.$$

Substituíndo as expresións anteriores

$$(e^a - 1)\left(\frac{1}{\lambda} - \frac{L}{e^a - 1}\right) + L = \frac{e^a - 1}{\lambda}.$$

Como $\lambda = \frac{N}{L_G}$ e $a = \frac{N \cdot L}{L_G}$ obtemos

$$\textbf{Tamaño medio dun contig} = L\frac{e^a - 1}{a}.$$

Capítulo 5

Caso práctico

5.1. Introdución

Neste capítulo poremos en práctica os conceptos sobre o ensamblado de secuencias estudados anteriormente. Para iso, realizaremos un exemplo práctico de ensamblado de lecturas, proceso que forma parte da secuenciación *shotgun*.

Esta técnica permite secuenciar longas cadeas de ADN e incluso xenomas completos. A secuenciación *shotgun* baséase na fragmentación do ADN en pequenos anacos que posteriormente son secuenciados, obtendo como resultado lecturas. Estas ensámblanse computacionalmente grazas á existencia de solapamento entre elas dando lugar aos contigs que agrupados forman unha representación do xenoma.

Unha das empresas especializadas nas técnicas de secuenciación mediante as cales se obteñen as lecturas é **Illumina** (`https://www.illumina.com`). Illumina é unha compañía estadounidense de biotecnoloxía que proporciona as ferramentas e servizos para a secuenciación do ADN, como kits de laboratorio ou equipos de secuenciación. Illumina emprega tecnoloxías de secuenciación de nova xeración (**NGS**, *Next Generation Sequencing*), técnicas que permiten determinar unha secuencia de ADN de forma moi rápida e precisa.

Entre as bases de datos sobre a secuenciación de ADN máis completas destaca o **NCBI** (*National Center for Biotechnology Information*), (`https://www.ncbi.nlm.nih.gov`), que forma parte da NLM (*National Library of Medicine*), unha rama do NIH (*National Institutes of Health*) dos Estados Unidos. O NCBI almacena diferentes bases de datos relacionadas co xenoma e a biomedicina, así como artigos sobre estudios científicos relacionados coa bioinformática e investigación biomédica.

Este centro de información emprégase frecuentemente no mundo da biotecnoloxía, xa que está en

constante actualización ao ritmo dos avances científicos. Ademais é de balde e está dispoñible en liña, o que facilita o seu acceso a toda a comunidade científica, impulsando novos descubrimentos científicos.

Dentro do NCBI atópase o **SRA** (*Sequence Read Archive*) que é o maior repositorio público de datos de secuenciación de alto rendemento. Esta base de datos forma parte do INSDC (*International Nucleotide Secuence Database Collaboration*) polo que comparte datos co EBI (*European Bioinformatics Institute*) e co DDBJ (*DNA Database of Japan*). Grazas a esta colaboración internacional, no SRA podemos atopar gran cantidade de datos de secuenciación sen procesar, así como información de aliñamento que facilita a reconstrución de secuencias de ADN propulsando novos descubrimentos mediante o análise destas secuencias.

5.2. Ensamblado dun xenoma bacteriano

Nesta sección realizaremos o ensamblado de lecturas obtidas tras a secuenciación dun xenoma bacteriano.

As bacterias son microorganismos unicelulares procariotas. Os organismos procariotas teñen un xenoma máis pequeno e sinxelo que os organismos eucariotas. Este está formado por unha soa molécula de ADN que ronda entre un e cinco millóns de nucleótidos, varía segundo a especie. Isto fai que traballar na secuenciación do seu xenoma sexa máis sinxelo e menos custoso computacionalmente que traballar co xenoma doutros organismos.

Por outro lado, o xenoma dos organismos eucariotas está composto por múltiples moléculas de ADN e ronda os cinco mil millóns de nucleótidos. Isto retarda o proceso de ensamblado chegando a ser inasumible para un ordenador convencional.

Ademais, engadindo que as bacterias están presentes en todo o noso entorno, existen moitos máis datos do xenoma das bacterias que doutros organismos. En efecto, na base de datos do NCBI podemos atopar aproximadamente 52 120 datos sobre o xenoma de organismos eucariotas fonte aos 2 690 000 datos sobre o xenoma das bacterias.

Estas características fixeron que se escollese o xenoma bacteriano para representar ensamblado dun xenoma, sendo posible sen un equipo avanzado.

En concreto o organismo escollido é a bacteria *Bacillus vanillea* SMT-24. *Bacillus* é un xénero de bacteria que ten forma de bastón (bacilos). *Vanillea* é unha especie de bacterias dentro deste xénero. Mentres que SMT-24 é o indicador que se emprega nos laboratorios para identificar a cepa, é dicir, para identificar unha variante dentro dunha especie.

O tipo de ensamblado que realizaremos denomínase **ensamblado de novo**, é dicir, ensambla-

remos as lecturas sen coñecer o xenoma orixinal. Este tipo de ensamblado de secuencias de ADN permite reconstruír secuencias xenómicas sin a necesidade de ter unha secuencia modelo empregando datos de secuenciación de nova xeración.

Para levar a cabo o ensamblado de lecturas empregaremos un software de ensamblado de secuencias xunto cun arquivo que contén as lecturas obtidas como resultado do proceso de secuenciación.

O primeiro paso para ensamblar é a obtención dos datos cos que imos a traballar, o ficheiro de lecturas sen ensamblar. Para iso, descargamos do SRA un arquivo (`https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR33847918&display=download`) que contén as lecturas obtidas coa tecnoloxía Illumina. Este ficheiro de texto sen formato (coa extensión ".fastq" ou ".fasta") contén a información necesaria para determinar os distintos fragmentos de ADN en estudo.

No SRA tamén podemos atopar moita outra información acerca destes datos tal e como vemos na Figura 5.1, (`https://www.ncbi.nlm.nih.gov/sra/?term=SRX29068242`). Aquí atopamos a información relacionada co organismo ao que pertence o xenoma, datos sobre o proceso de obtención das lecturas e tamén a data de publicación do ficheiro.

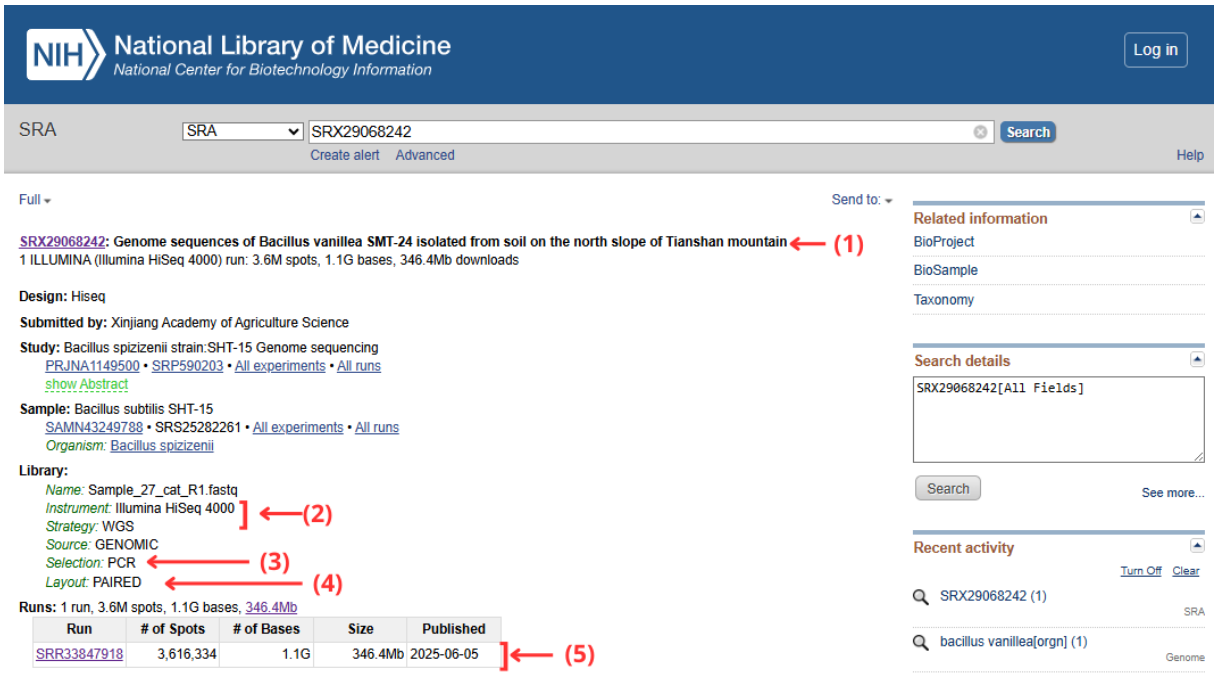

Figura 5.1: Datos sobre o ficheiro coas lecturas sen ensamblar.

Analizando a Figura 5.1 vemos que en (1) indícase que o xenoma pertence á bacteria *Bacillus vanillea* SMT-24 illada no chan na ladeira norte da montaña Tian Shan en Asia central.

En (2) observamos que o ficheiro foi xerado empregado o equipo Illumina HiSeq 4000. Ademais

as siglas WGS (*Whole Genome Sequencing*) indícanos que todo o xenoma foi secuenciado, sen centrarse nunha parte en concreto.

En (3) vemos que se empregou unha PCR (Reacción en Cadea da Polimerasa). Esta técnica ten como obxectivo obter un gran número de copias a partir dun fragmento de ADN, polo que se emprega habitualmente para obter copias dos fragmentos de ADN que sen van secuenciar e así mellorar fiabilidade dos resultados.

En (4) recóllese que se fixo unha secuenciación *paired-end*. Vexamos que significa isto, pero primeiro precisamos introducir o concepto de spot.

Chamamos **spot** a cada fragmento de ADN lido polo equipo de secuenciación, pero que non necesariamente ten que ser unha lectura. Así, cada spot pode estar formado por unha ou dúas lecturas.

Cando o spot está formado por unha soa lectura falamos de secuenciación *single-read*, quere dicir que o fragmento lese por un único extremo. Mentres que, cando un spot está formado por dúas lecturas falamos de secuenciación *paired-end*, quere dicir cada fragmento de ADN lese dende ambos extremos producindo así dúas lecturas, tal e como podemos ver na Figura 5.2.

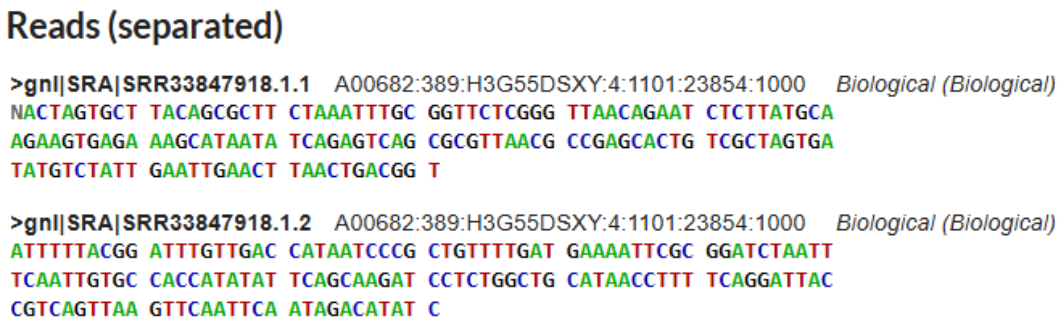

Figura 5.2: Exemplo de spot formado por dúas lecturas.

Observando a Figura 5.2 notamos que hai letras N que non se corresponden con ningún tipo de nucleótido (A, C, G, T). Isto indica que non se recoñeceu o nucleótido que ocupa ese lugar. Por outro lado a etiqueta *Biological* indícanos que esta é a estrutura da lectura tal e como se encontra na realidade, xa que este spot tamén pode representarse como unha única lectura.

En (5) vemos que o arquivo publicado o 5 de xuño de 2025 contén 3 616 334 spots. Cada spot contén dúas lecturas, cada unha de lonxitude 151 nucleótidos.

Por outro lado, o xenoma de dita bacteria ten unha lonxitude de 3 800 000 nucleótidos. Polo tanto a cobertura, é dicir, o número medio de veces que se secuencia cada nucleótido do xenoma

orixinal é

$$\frac{3\ 616\ 334 \cdot 2 \cdot 151}{3\ 800\ 000} \approx 287.$$

É dicir temos unha cobertura de $287X$.

Unha vez obtidos e interpretados os datos precisamos dunha ferramenta que nos permita ensamblalos e así obter os contigs.

O **SPAdes** (St. Petersburg genome assembler) [7] é un software de ensamblado e análise de datos de secuenciación dispoñible de forma libre na plataforma GitHub (`https://github.com/ablab/spades`). Este programa esta especialmente deseñado para datos de secuenciación de Illumina.

A linguaxe de programación maioritaria de SPAdes é C++ aínda que tamén algo de C e Python. No obstante, non traballaremos co código fonte, senón que usaremos os binarios, é dicir, cos arquivos executables xerados compilando o código fonte na terminal.

Empregamos a versión de SPAdes 4.2.0. que está dispoñible para os sistemas operativos de Linux de 64 bits e Mac.

Para empregar os binarios só temos que descargar e descomprimir o arquivo. Para iso imos á terminal, cambiamos ao directorio onde queremos que se instale e executamos os dous seguintes comandos:

```
wget https://github.com/ablab/spades/releases/download/v4.2.0/SPAdes-4.2.0-
Linux.tar.gz
tar -xzf SPAdes-4.2.0-Linux.tar.gz
```

A continuación verificamos que a instalación está correcta executando o seguinte comando na terminal:

```
bin/spades.py -test
```

Se todo está correcto debemos de atopar a seguinte mensaxe na saída da terminal:

```
========= TEST PASSED CORRECTLY.
```

Agora xa está todo preparado para ensamblar.

Executamos o seguinte comando na terminal

```
bin/spades.py -12 SRR33847918.fastq -o /home/celiadr145/Downloads/SPAdes-
4.2.0-Linux/saida -only-assembler
```

Figura 5.3: Pantalla da terminal tras executar o ensamblado de secuencias.

Tal e como vemos na Figura 5.3, ao executar este comando a saída por pantalla describe cada unha das ordes tal e como indicamos a continuación:

- `--12 SRR33847918.fastq` indica que as dúas lecturas de cada spot están no mesmo arquivo SRR33847918.fastq. No caso contrario, se as lecturas de cada spot estivesen separadas en dous arquivos escribiríamos -1 nomearquivo1.fastq -2 nomearquivo2.fastq.
- `--only-assembler` indica a SPAdes que omita a parte de corrección de erros e pase ao ensamblado.
- `-o /home/celiadr145/Downloads/SPAdes-4.2.0-Linux/saida` indica que queremos que cree unha cartafol chamado "saida" nese enderezo no que garde os resultados da execución.

Despois de 15 minutos e 42 segundos remata a execución co aviso por pantalla

```
Assembling time: 0 hours 15 minutes 42 seconds
...
===== Assembling finished.
...
* Assembled contigs are in /home/celiadr145/Downloads/SPAdes-4.2.0-Linux
/saida/contigs.fasta
```

Como podemos observar, crea un ficheiro dentro da cartafol "saida" chamado contigs.fasta que podemos abrir co editor de texto.

Cada contig vén marcado co símbolo ">", así facendo unha busca podemos ver que obtivemos 1811 contigs, tal e como vemos na Figura 5.4.

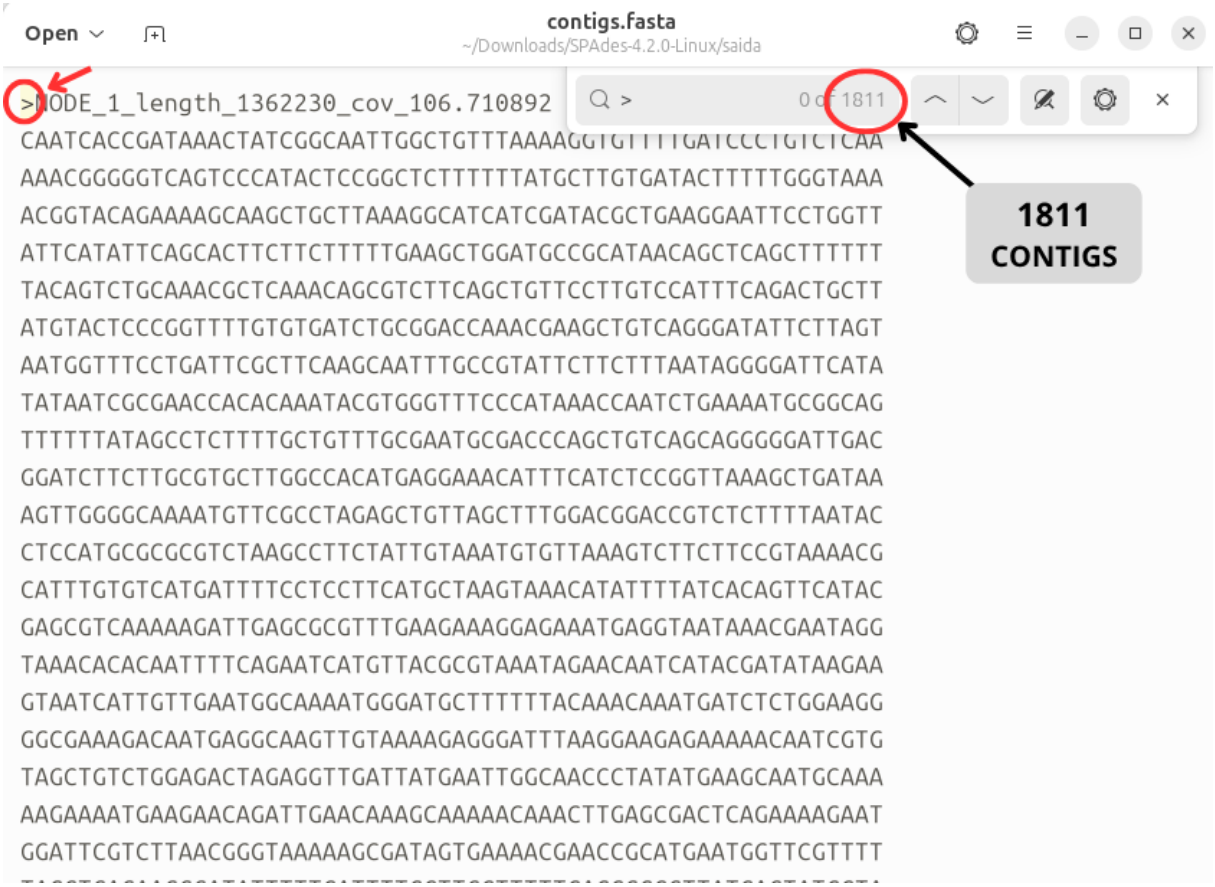

Figura 5.4: Visualización do ficheiro contigs.fasta.

Con este caso práctico rematamos o noso traballo. Vimos que grazas á estadística é posible resolver diversas cuestións relacionadas co ensamblado do ADN. O estudo destas cuestións permite o desenvolvemento de técnicas de ensamblado cada vez máis avanzadas, co obxectivo de conseguir unha representación mediante contigs o máis semellante posible ao xenoma orixinal. Este avance permite un maior coñecemento do material xenético, de gran importancia, xa que é o que determina e controla o funcionamento de todos os seres vivos.

Bibliografía

[1] Aguado, F., Gago, F., Ladra, M., Pérez, G., Vidal, C. e Vieites, A. M. (2018). *Problemas resueltos de combinatoria. Laboratorio con Sagemath*, 1st ed., Ediciones Paraninfo.

[2] Cooper, G. M. e Hausman, R. E. (2017). *La Célula*, 7th ed., Marbán.

[3] Ewens, W. J. e Grant, G. R. (2005). *Statistical Methods in Bioinformatics*, 2nd ed., Springer.

[4] Nores, M. L. (2021). *Un recorrido por las distribuciones de probabilidad*, Revista de Educación Matemática, **36**, 7–43.

[5] Rohatgi, V. K. e Saleh, A. K. Md. E. (2000). *An Introduction to Probability and Statistics*, 2nd ed., John Wiley & Sons.

[6] Ross, S. M. (1996). *Stochastic Processes*, 2nd ed., John Wiley & Sons.

[7] Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. e Korobeynikov, A. (2020). *Using SPAdes de novo assembler*, Current Protocols in Bioinformatics, **70**, e102.

[8] Vélez Ibarrola, R. (2019). *Cálculo de probabilidades 2*, UNED.