

# Contenido

## **PRIMERA PARTE**

Secuencias génicas y de proteínas	31
<b>Capítulo 1. ¿Cómo comenzar a trabajar?</b>	33
1.1. Introducción	35
1.2. ¿Qué debe esperar el lector?	36
1.3. Breve reseña histórica	37
1.4. La internet	37
1.5. ¿Qué es un dato en biología?	40
1.6. Algunos tópicos importantes relacionados con la bioinformática	41
1.7. ¿Qué hacer?	46
<b>Capítulo 2. Bases de datos</b>	49
2.1. Introducción	51
2.2. ¿Qué son las herramientas o “Tools” en Bioinformática?	52
2.3. ¿Qué hacer? Parte 1	54
2.4. ¿Cómo escoger una base de datos?	55
2.5. Las bases de datos más importantes	55
2.6. ¿Qué hacer? Parte 2	57
<b>Capítulo 3. Secuencias</b>	59
3.1. Introducción	61
3.2. Estructura de una secuencia	62
3.3. ¿Qué es el formato de una secuencia?	64
3.4. ¿Cuáles son los formatos más utilizados?	64

3.5.	¿Cómo ver una secuencia?	70
3.6.	¿Qué hacer? Parte 1	71
3.7.	¿Qué hacer? Parte 2	73
<b>Capítulo 4. Uso de PubMed y MEDLINE</b>		<b>75</b>
4.1.	Introducción	77
4.2.	¿Cuál es la diferencia entre PubMed y MEDLINE?	78
4.3.	¿Cómo acceder a PubMed?	78
4.4.	¿Qué hacer? Parte 1	79
4.5.	Otras estrategias de búsqueda en PubMed	80
	4.5.1. Búsqueda avanzada	80
4.6.	¿Qué hacer? Parte 2	82
<b>Capítulo 5. Trabajando con secuencias de proteínas</b>		<b>83</b>
5.1.	¿Qué información se obtiene de una secuencia de proteínas?	85
5.2.	Explorando otras formas de obtener información de la secuencia de una proteína	88
5.3.	Procariontes vs. eucariotes	90
5.4.	¿Qué hacer?	91
<b>Capítulo 6. Trabajando con secuencias de ADN</b>		<b>93</b>
6.1.	¿Qué información se encuentra en una secuencia de ADN?	95
6.2.	¿Cómo buscar una secuencia de adn?	96
6.3.	¿Qué hacer? Parte 1	97
6.4.	Otras bases de datos con secuencias de ADN	99
6.5.	Búsqueda avanzada de secuencias de ADN	100
6.6.	¿Qué otros tipos de ADN hay?	102
6.7.	¿Qué hacer? Parte 2	103
6.8.	Limpieza de una secuencia	103
6.9.	Mapas de restricción de una secuencia génica	105
6.10.	Análisis del contenido de una secuencia génica	108
6.11.	¿Cómo localizar palabras en secuencias de ADN?	108

6.12. ¿Qué hacer? Parte 3.	110
6.13. Localización de secuencias tipo “Repeats” en el genoma de los organismos	110
<b>Capítulo 7. Mapeo genético y Bases de Datos de Mapeo</b>	113
7.1 ¿Qué es el mapeo genético y cuál es su importancia actual?	117
7.2. ¿Cómo acceder a las bases de datos de mapeo génico?	118
7.3. ¿Qué hacer?	119
7.4. Explorando otras bases de datos genómicos	121
7.5. Ensamblajes de genomas	122
<b>SEGUNDA PARTE</b>	
<b>Explorando las secuencias de ácidos nucleicos y proteínas</b>	125
<b>Capítulo 8. ¿Cómo buscar similitud entre las secuencias?</b>	127
8.1. ¿Qué es la herramienta BLAST?	129
8.2. Herramientas básicas de BLAST	130
8.2.1. “Nucleotide BLAST”	130
8.2.2. “Protein BLAST” (Blastp)	134
8.2.3. Otras herramientas de BLAST	135
8.3 ¿Qué hacer?	136
<b>Capítulo 9. Exploración profunda de base de datos de proteínas</b>	139
9.1. Información de proteínas que se encuentra en las bases de datos	141
9.2. ¿Qué hacer? Parte 1	143
9.3. El marco abierto de lectura de un gen que se traduce	143
9.4. ¿Qué hacer? Parte 2	144
9.5. Información sobre la función de las proteínas	145
9.6. ¿Qué Hacer? Parte 3	147
9.7. La estructura de una proteína y otros datos importantes	149
9.8. ¿Qué hacer? Parte 4	153

<b>Capítulo 10. Otras herramientas útiles en el estudio de las proteínas</b>	155
10.1. Herramientas para la caracterización de una proteína	157
10.2. Parámetros de caracterización fisicoquímica de una proteína	161
10.3. Otras herramientas para caracterizar la secuencia de una proteína	162
10.4. ¿Qué hacer?	163
<b>Capítulo 11. Análisis funcional de una proteína</b>	165
11.1. El carácter hidrofóbico de un péptido	167
11.2. ¿Cómo mejorar el análisis de la estructura de una proteína?	171
11.3. Biología evolutiva de secuencias de proteínas	176
11.4. ¿Cómo se detectan nuevos dominios en una secuencia de una proteína?	177
11.5. ¿Qué hacer?	178
<b>Capítulo 12. Alineamiento de secuencias de ADN y proteínas</b>	181
12.1. ¿Para qué se alinean las secuencias?	183
12.2. ¿Cuál es significado biológico de la similitud?	184
12.3. ¿Cómo se calcula la similitud entre dos secuencias?	185
12.3.1. Alineamientos globales de secuencias	185
12.3.2. Análisis de dos secuencias por alineamiento manual	191
12.3.3. Alineamientos locales de secuencias	194
12.4. ¿Qué hacer?	199
<b>Capítulo 13. Alineamiento de múltiples secuencias</b>	201
13.1. ¿Qué es un alineamiento múltiple de secuencias y para qué se utiliza?	203
13.2. Formato de entrada de datos para alineamiento múltiple	206
13.3. ¿Cómo se lee un alineamiento múltiple?	208
13.4. Clustal	211
13.5. T-COFFEE	213

13.6. EMBOSS Cons	216
13.7. MUSCLE	218
13.8. ¿Qué hacer?	219
<b>Capítulo 14. Revisión y edición de secuencias para su publicación</b>	<b>221</b>
14.1. ¿Para qué se publican las secuencias de nucleótidos o de proteínas?	223
14.2. Curación de una secuencia	224
14.3. Edición de alineamientos múltiples de secuencias de ADN	225
14.4. ¿Cómo someter una secuencia a las bases de datos?	229
14.5. Edición de múltiples secuencias de proteínas	230
Referencias	235
Índice analítico	255

# Índice de tablas y figuras

# Figuras

<b>Figura 1.1.1.</b> Relevancia actual del manejo de datos	37
<b>Figura 1.4.1.</b> Áreas biológicas e informáticas que alimentan la bioinformática	40
<b>Figura 1.4.2.</b> Modelo de la estructura del ADN, (Freeman, 2000)	41
<b>Figura 1.6.1.</b> El ADN y el ARN son polímeros de nucleótidos y el ADN, a diferencia del ARN, no tiene la base uracilo aparte de otras diferencias químicas (Alberts <i>et al.</i> , 2002)	44
<b>Figura 1.6.2.</b> La relación de las especies moleculares de las células vivas (ADN, ARN, proteínas y metabolitos) con las diferentes áreas de estudio	45
<b>Figura 1.6.3.</b> El código genético (Wikimedia commons, 2017a)	46
<b>Figura 2.2.1.</b> La clasificación, el manejo y uso de los datos es el objeto de la Bioinformática.	54
<b>Figura 3.1.1.</b> Relaciones entre la secuencia de ADN, el ARN y la secuencia de aminoácidos en un péptido	63
<b>Figura 3.4.1.</b> Comparación de una secuencia de aminoácidos con su correspondiente secuencia de nucleótidos	71
<b>Figura 3.5.1.</b> Cromatograma de una secuencia de ADN obtenida por el método de Sanger	72
<b>Figura 3.6.1.</b> Estructura tridimensional de la ribonucleasa pancreática bovina (Cleanpng, s.f.)	74
<b>Figura 4.2.1.</b> Las diferentes opciones de búsqueda en “Home” de PubMed.	80
<b>Figura 4.5.1.</b> Portada de algunas de las principales publicaciones científicas a nivel internacional	83
<b>Figura 5.1.1.</b> “Home” de ExPasy donde están las entradas a las bases de datos y las principales herramientas de trabajo con secuencias	87
<b>Figura 5.1.2.</b> Resultados de la búsqueda en la base de datos de ExPASy sobre el tema de la proteína P53	88
<b>Figura 5.1.3.</b> Herramientas que se encuentran en el portal neXtProt de UniProt	90

<b>Figura 5.2.1.</b> Presentación de la secuencia de aminoácidos de una proteína, en formato plano de texto	91
<b>Figura 5.2.2.</b> Salida de la información gráfica de la secuencia del gen P53	92
<b>Figura 5.4.1.</b> “Homepage” del sitio más popular de información de proteínas en EEUU, en el NCBI	94
<b>Figura 6.2.1.</b> Homepage de Hüge Gene Nomenclature Committee	99
<b>Figura 6.2.2.</b> “Homepage” de la herramienta “Nucleotide” del NCBI para análisis de secuencias de ADN	99
<b>Figura 6.3.1.</b> Falla en la búsqueda de secuencias en el portal “Nucleotide” del NCBI	101
<b>Figura 6.5.1.</b> Links que enlazan diferentes bases de datos a una búsqueda específica en el portal de UniProt	103
<b>Figura 6.8.1.</b> Portal de la herramienta VecScreen que permite localizar secuencias de vectores que contaminan una secuencia	106
<b>Figura 6.8.2.</b> Ejemplo de la salida del programa VecScreen	106
<b>Figura 6.9.1.</b> “Home” del portal de REBASE donde se puede consultar los sitios de restricción de una enzima por su nombre	107
<b>Figura 6.9.2.</b> Sitio de la herramienta WatCut para obtener mapas de restricción de secuencias génicas	108
<b>Figura 6.9.3.</b> “Homepage” de la herramienta “Restriction Map” del portal “Sequence Manipulation Suite”	109
<b>Figura 6.9.4.</b> Resultado del análisis de los posibles sitios de restricción de una secuencia génica	109
<b>Figura 6.11.1.</b> Portal del programa Wordcount para detectar palabras en una secuencia de nucleótidos y donde se puede determinar el tamaño de las palabras	111
<b>Figura 6.11.2.</b> “Homepage” de la base de datos de uso de tripletes de secuencias de genes de diferentes organismos (Codon Usage Database), organizadas por el nombre de la especie	112
<b>Figura 6.13.1.</b> Portal de la herramienta “Tanden Repeats Finder” de las University of Boston (Massachusetts)	113
<b>Figura 7.1.1.</b> “Homepage” del recurso mapeo genético de NIH	117
<b>Figura 7.1.2.</b> Homepage de Illumina®	118

<b>Figura 7.1.3.</b> Relación entre la secuencia genómica, el mapa físico (posición física de un gen en un cromosoma), el mapa citogenético y el mapa génico de un cromosoma. Fuente: NIH (EEUU)	119
<b>Figura 7.2.1.</b> “Homepage” del Instituto de Ciencias Genómicas de la Universidad de California	121
<b>Figura 7.3.1.</b> Barra de herramientas del explorador de cromosomas en NCBI.	122
<b>Figura 7.4.1.</b> Portal de NCBI especializado en información génica de microorganismos (Bacterias), Virus y organelos como mitocondrias y plastidios	123
<b>Figura 7.5.1.</b> Portal del proyecto “Ensemble Genomes” de European Bioinformatics Institute del EMBL-EBI (UK)	125
<b>Figura 7.5.2.</b> Portal del sitio de ensamblaje del genoma humano de EMBL-EBI	125
<b>Figura 8.2.1.1.</b> Ventana de acceso para la información de búsqueda de secuencias de nucleótidos utilizando la herramienta Blastn	132
<b>Figura 8.2.1.2.</b> Opciones de búsqueda de Blastn	133
<b>Figura 8.2.1.3.</b> Primera parte del resultado de búsqueda con la herramienta BLAST	134
<b>Figura 8.2.1.4.</b> Descripción de los resultados de la búsqueda en Blastn	134
<b>Figura 8.2.1.5.</b> Comparación de dos secuencias génicas producto de la búsqueda de secuencias similares con Blastn	135
<b>Figura 8.2.2.1.</b> Fragmento de la comparación de dos secuencias de proteínas usando Blastp	136
<b>Figura 8.2.3.1.</b> Resultado de la búsqueda de una secuencia de nucleótidos en bases de datos de proteínas utilizando la herramienta Blastx	138
<b>Figura 9.1.1.</b> Base de datos con información sobre proteínas fosforiladas dbPAF (database of Phospho-sites in Animals and Fungi). Fuente: GPS, 2021)	144
<b>Figura 9.3.1.</b> “Homepage” de la herramienta de detección de ORFs en secuencias génicas o de mRNA del NCBI	146
<b>Figura 9.5.1.</b> Identificación de una proteína en la base de datos UniProt.	147
<b>Figura 9.5.2.</b> Presentación de la información de la función de una proteína en el portal UniProt	148
<b>Figura 9.5.3.</b> Portal de InterPro donde se puede encontrar información sobre la función y la estructura de una proteína particular	149
<b>Figura 9.6.1.</b> Bases de datos en el portal InterPro	150

<b>Figura 9.7.1.</b> Información general sobre una proteína en el portal de estructura de proteínas en RCSB PDB	152
<b>Figura 9.7.2.</b> Datos sobre el método de estudio de la estructura de una proteína e información técnica sobre las condiciones de la técnica utilizada	152
<b>Figura 9.7.3.</b> Información bibliográfica que se presenta en el portal del PDB sobre una proteína particular	153
<b>Figura 9.7.4.</b> Estructura tridimensional de la proteína 2W6V	153
<b>Figura 9.7.5.</b> Información completa sobre las diferentes características estructurales de la proteína en estudio	154
<b>Figura 9.7.6.</b> Información y link sobre el gen correspondiente a la proteína en estudio	155
<b>Figura 10.1.1.</b> Ventana de entrada de la información en la herramienta FindMod del portal de ExPasy	160
<b>Figura 10.1.2.</b> Entrada de información de la herramienta FindPep de ExPASy	160
<b>Figura 10.1.3.</b> Entrada a la herramienta Profound que permite reconocer péptidos para realizar un mapa de una proteína probable	162
<b>Figura 10.1.4.</b> “Homepage” de la herramienta ProteinProspector de la University of California para la construcción de mapas de péptidos, producto de la técnica de espectrometría de masas	162
<b>Figura 10.2.1.</b> Entrada de la herramienta ProtParam de ExPASy que nos permite encontrar diferentes parámetros fisicoquímicos, partiendo de la secuencia de una proteína	163
<b>Figura 10.2.2.</b> “Home” de la herramienta Compute pi/Mw para calcular el punto isoeléctrico y la masa molecular de la proteína, cuya secuencia entramos	164
<b>Figura 10.3.1.</b> Página principal de la herramienta PeptideCutter de ExPASy donde se puede conocer si una proteína tiene sitios que puedan ser atacados por peptidasas o por químicos	165
<b>Figura 11.1.1.</b> Entrada de la herramienta ProtScale de ExPASy	170
<b>Figura 11.1.2.</b> Listado de opciones para análisis de la estructura secundaria de una proteína con la herramienta ProtScale	171
<b>Figura 11.1.3.</b> Gráfica sobre la distribución de aminoácidos hidrofóbicos en la secuencia de una proteína	172

<b>Figura 11.1.4.</b> “Home” de la herramienta TMHMM de DTU Health Tech para análisis de estructuras hidrofóbicas en proteínas	173
<b>Figura 11.2.1.</b> “Homepage” de la herramienta “Coils” de ExPASy	174
<b>Figura 11.2.2.</b> “Homepage” de PROSITE de ExPASy	175
<b>Figura 11.2.3.</b> Entrada a la herramienta InterProScan, útil para determinar si una secuencia de una proteína está relacionada con una familia de proteínas preexistentes	177
<b>Figura 11.2.4.</b> Forma en que se presenta el resultado de una búsqueda de familias de proteínas en la herramienta InterProScan. Para acceder a los datos del resultado se debe hacer clic en la referencia	177
<b>Figura 11.3.1.</b> Entrada a las herramientas de la CDD (Conserved Domain Database) en el NCBI	178
<b>Figura 11.3.2.</b> <i>Software</i> que permite la búsqueda de dominios conservados en una secuencia de una proteína o en una secuencia de nucleótidos de un gen que codifica para una proteína	179
<b>Figura 12.2.1.</b> Similaridad encontradas en el gen de la $\beta$ -Tubulina en diferentes especies del género <i>Fusarium</i> , utilizando el programa CLC WorkBench®	186
<b>Figura 12.2.3.2.</b> Herramienta Dotlet JS beta	195
<b>Figura 12.3.1.1.</b> Entrada de la herramienta BLAST en ExPASy del SIB	189
<b>Figura 12.3.1.2.</b> Ventana para escoger la base de datos que se usará para realizar la búsqueda (BLAST) en el sitio Blast “Form”	189
<b>Figura 12.3.1.3.</b> Resultado de la búsqueda BLAST “Form” de una secuencia de proteína	190
<b>Figura 12.3.1.4.</b> Información que el programa Blast Form da sobre la comparación de dos secuencias	191
<b>Figura 12.3.1.5.</b> Entrada de datos para el programa Needle de EMBL-EBI.	192
<b>Figura 12.3.1.6.</b> Entrada de la herramienta para comparar secuencias EMBOSS “Stretcher” de EMBL-EBI	192
<b>Figura 12.3.2.1.</b> Entrada de la herramienta Dotlet JS beta de SiB	194
<b>Figura 12.3.2.3.</b> Densidad de los puntos y densidad del fondo en el programa Dotlet JS beta	196
<b>Figura 12.3.3.1.</b> Herramientas básicas del BLAST de NCBI. Entrada al BLAST de nucleótidos y al BLAST de proteínas	197

<b>Figura 12.3.3.2.</b> Entrada a la página de la herramienta LALIGN de ExPASy para realizar alineamientos locales de dos secuencias	198
<b>Figura 12.3.3.3.</b> Alineamiento local	198
<b>Figura 12.3.3.4.</b> Introducción de <i>gaps</i> en cuatro secuencias que se están alineando	199
<b>Figura 12.3.3.5.</b> Entrada de la herramienta EMBOSS Water, para el alineamiento local de dos secuencias	200
<b>Figura 12.3.3.6.</b> Entrada a la herramienta BLAST de UniProt	200
<b>Figura 12.4.1.</b> Ejemplo de región homóloga con alineamiento de la secuencia	202
<b>Figura 13.1.1.</b> Árbol filogenético que relaciona 15 especies del género <i>Fusarium</i> utilizando secuencias de genes útiles para análisis taxonómico	206
<b>Figura 13.1.2.</b> Resultado del alineamiento de nueve secuencias de aminoácidos de la proteína Citocromo C oxidasa de diferentes especies de lamprea	207
<b>Figura 13.2.1.</b> Home page de la herramienta EMBOSS Segret que permite inter-convertir formatos	209
<b>Figura 13.3.1.</b> Secuencias nucleotídicas no alineadas en la parte superior y las mismas secuencias, pero alineadas utilizando un software para tal fin en la parte inferior	210
<b>Figura 13.4.1.</b> “Homepage” de la herramienta Clustal. Aquí se pueden obtener las versiones antiguas y la última versión de la herramienta de alineamiento múltiple	213
<b>Figura 13.4.2.</b> “Home” de la herramienta Clustal Omega en el portal de EMBL-EBI	214
<b>Figura 13.4.3.</b> Resultados presentados por la herramienta Clustal Omega de un alineamiento de cinco secuencias de un gen mitocondrial (FM1) de diferentes especies de <i>Fusarium</i>	215
<b>Figura 13.5.1.</b> Página principal de la herramienta T-Coffee en la plataforma EMBL-EBI	216
<b>Figura 13.5.2.</b> Alineamiento de cinco secuencias de el gen FM1 de especies de <i>Fusarium</i> obtenidas con la herramienta T-Coffee	217
<b>Figura 13.5.3.</b> Árboles filogenéticos obtenidos con las herramientas Clustal (arriba) y T-Coffee abajo	218

<b>Figura 13.6.1.</b> Portal de la herramienta <i>EMBOSS Cons</i> , que permite obtener la mejor secuencia consenso entre un grupo de secuencias alienadas apropiadamente	219
<b>Figura 13.6.2.</b> Resultado en la obtención de una secuencia consenso de un alineamiento múltiple utilizando la herramienta <i>EMBOSS Cons</i> de EMBL-EBI	219
<b>Figura 13.6.3.</b> Alineamiento múltiple utilizando el <i>software</i> de <i>CLC Main Workbench</i> <sup>®</sup>	220
<b>Figura 14.4.1.</b> Página principal del tutorial para subir una secuencia genómica a la base de datos de nucleótidos del <i>NCBI</i>	232
<b>Figura 14.3.1.</b> Alineamiento múltiple de secuencias de <i>ADN</i> utilizando la herramienta <i>ChromasPro</i> <sup>®</sup> de la empresa <i>Technelysium</i> .	227
<b>Figura 14.3.2.</b> Página de entrada de la herramienta <i>EMBOSS Seqret</i> de EMBL-EBI	228
<b>Figura 14.3.3.</b> “Homepage” de la herramienta “Reverse Complement” de <i>GenScript Corporation</i> para obtener el complemento reverso de una cadena de <i>ADN</i>	229
<b>Figura 14.3.4.</b> “Homepage” de la herramienta <i>MView</i> para hacer alineamientos múltiples de secuencias de <i>ADN</i> o de proteínas	230
<b>Figura 14.3.5.</b> Eliminación de colas para obtener el consenso útil para enviar a las bases de datos	231
<b>Figura 14.5.1.</b> Instrucciones para crear el archivo <i>FASTA</i> de la secuencia que se quiere someter	233
<b>Figura 14.5.2.</b> Portal de la herramienta <i>Jalview</i> para alinear secuencias de <i>ADN</i> o de proteínas, además de contar con otras utilidades como programas para hacer filogenia	234
<b>Figura 14.5.3.</b> Herramienta <i>Jalview</i> para editar secuencias de proteínas.	235
<b>Figura 14.5.4.</b> Instrucciones iniciales para someter secuencias de proteínas en la base de datos de <i>UniProt</i>	235
<b>Figura 14.5.5.</b> “Homepage” del portal de <i>UniProtKB</i> de EMBL-RBI para cargar una secuencia de una proteína en la base de datos	236

# Tablas

<b>Tabla 1.6.1.</b> Códigos de los veinte aminoácidos más comunes en las proteínas que mantienen su derivación del nombre en inglés	48
<b>Tabla 2.5.1.</b> Otras bases de datos complementarias a las primarias y secundarias	59
<b>Tabla 12.3.1.</b> Ejemplo de consenso en el apareamiento	188
<b>Tabla 12.3.2.1.</b> Matriz de apareamiento de dos secuencias con diferencias en la información contenida en ellas	193
<b>Tabla 13.3.1.</b> Arreglo de múltiples secuencias cortas de nucleótidos para obtener la secuencia consenso	212