



EpiMolBio

Análisis de variabilidad genética
Manual de usuario v 1.0



**LABORATORIO DE
EPIDEMIOLOGÍA
MOLECULAR DEL VIH**
IRYCIS-Servicio de Microbiología
Hospital Ramón y Cajal





EpiMolBio

Análisis de la variabilidad genética

Licencia: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (número de registro 2305114294344)

Desarrollador: Roberto Reinosa

Colaboradora: Paloma Troyano-Hernáez

Coordinación y supervisión: África Holguín

Copyright: Roberto Reinosa Fernández y Fundación para la Investigación Biomédica del Hospital Universitario Ramón y Cajal (FIBioHRC)

INTRODUCCIÓN

Bienvenido a EpiMolBio, un programa bioinformático gratuito para el análisis de secuencias genéticas y proteicas. Este software permite analizar secuencias genéticas y obtener información sobre su estructura, función y evolución. Tanto si es un bioinformático experimentado como un biólogo principiante, este programa está diseñado para ser intuitivo y fácil de usar, con potentes herramientas y flujos de trabajo que pueden personalizarse para adaptarse a sus necesidades.

La versión de EpiMolBio 0.1.1 permite:

- Identificar mutaciones de resistencia en VIH-1 y VIH-2, su frecuencia y la conservación de las tres proteínas Pol del VIH: Proteasa, Retrotranscriptasa o Transcriptasa Inversa e Integrasa.
- Rastrear proteínas específicas dentro del genoma de SARS-CoV-2.
- Analizar la variabilidad genética de cualquier grupo de secuencias obteniendo información sobre su conservación, frecuencia de polimorfismos y mutaciones, generando secuencias consenso y tablas del índice de Wu-Kabat.
- Analizar la similitud de cualquier grupo de secuencias genéticas.
- Rastrear secuencias específicas de interés dentro de una o varias secuencias genéticas completas.
- Realizar alineamientos simples o múltiples de más de 100.000 secuencias.
- Realizar otras funciones variadas sobre las secuencias con la función Herramientas: desde la traducción de nucleótidos a aminoácidos hasta la edición de archivos.
- Programar funciones para automatizar su ejecución de forma encadenada.

EpiMolBio ha implementado múltiples funciones para el estudio de la variabilidad genética de dos de los patógenos que causan pandemias actualmente: el virus de la inmunodeficiencia humana (VIH), causante del SIDA, y el virus SARS-CoV-2, causante del COVID-19, patógenos responsables de más de 85 y 450 millones de infecciones en el mundo, respectivamente.

El programa está en continua mejora para adaptarse a las necesidades de nuevos proyectos de investigación. EpiMolBio se puede aplicar al estudio de la variabilidad genética de patógenos, y al análisis de marcadores genéticos asociados a enfermedades y de genes o proteínas de interés biomédico o biológico. Este programa se ha ideado como apoyo para la investigación de nuevas estrategias diagnósticas, pronósticas o terapéuticas de patógenos o enfermedades, y de estudios epidemiológicos de presencia de biomarcadores asociados a enfermedades o procesos biológicos de interés.

En este Manual de Usuario se detalla cómo llevar a cabo todas estas tareas. En la web www.epimolbio.com proporcionamos documentación completa, videotutoriales y asistencia al usuario para ayudarle a sacar el máximo partido del programa y alcanzar sus objetivos de investigación.

Este programa es multiplataforma, puede usarse en Windows , Mac, y Linux. A su vez es portable (se puede cortar y pegar en el escritorio o en una carpeta), por lo que **no requiere instalación**. Como requisito es necesario tener instalado Java 11 o superior, disponible en <https://www.java.com> de forma gratuita.

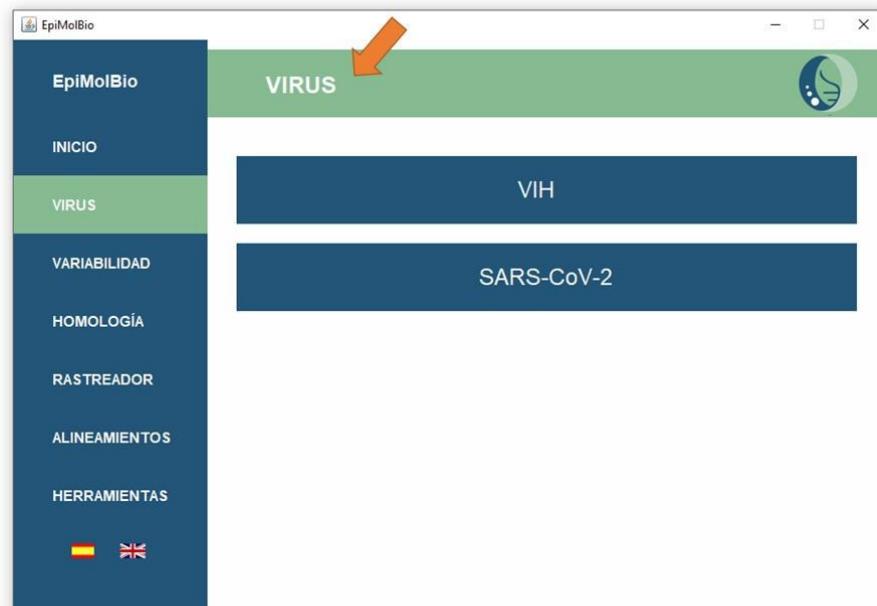
GENERALIDADES

La **interfaz** de EpiMolBio está diseñada para ser lo más sencilla e intuitiva posible, de manera que no es necesario poseer conocimientos previos de programación o bioinformática para poder utilizarlo.

A la izquierda se encuentra el **Menú** a través del cual se accede a las distintas funciones. Debajo están las opciones para seleccionar el **idioma** del programa (español o inglés).



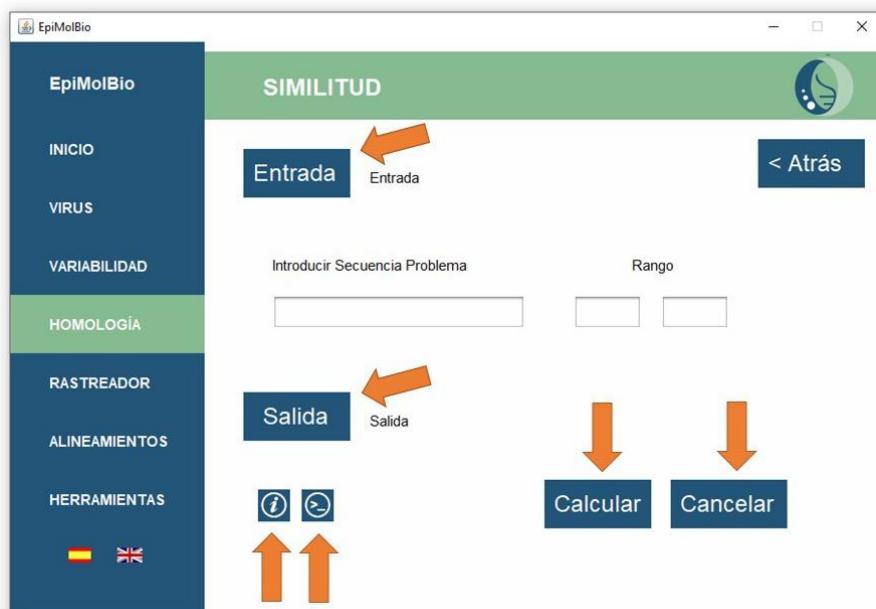
La cabecera indicará en todo momento cual es la última opción que se ha escogido.



El botón < Atrás permite retroceder al menú anterior.



En la mayoría de funciones se encontrarán los botones de **Entrada**, donde se escoge la carpeta que contiene las secuencias (en carpetas y/o en subcarpetas) a analizar; **Salida**, donde se escoge el nombre del archivo o carpeta con los resultados del análisis; **Calcular**, para llevar a cabo el análisis; **Cancelar**, para interrumpir un análisis en curso; **Botón de Información**, con un resumen de los pasos para llevar a cabo el análisis y los formatos de salida que permite cada función; y **Botón de programación de funciones**, que registra las opciones que se han escogido y puede copiarse como input de la herramienta Programar Funciones.



En la mayoría de funciones de EpiMolBio los **tipos de archivo de entrada** son archivos de secuencias del tipo **.fasta**, que bien pueden ubicarse dentro de una carpeta (generalmente) o en subcarpetas dentro de una carpeta cuando se requiere hacer ciertas funciones. Sólo en casos concretos se requerirá otro formato de archivo, como archivos de texto **.txt**.

Aunque algunas funciones están diseñadas para emplear archivos .fasta en aminoácidos, también pueden emplearse archivos en nucleótidos si previamente se emplea la herramienta Buscar y Reemplazar de Edición de Archivos sustituyendo las “N” (nucleótido desconocido) por “?” (aminoácido desconocido) para que estas se excluyan del análisis, como se explicará en detalle en la sección Herramientas de este manual.

La mayoría de **archivos de salida** resultantes de los análisis de EpiMolBio son de tipo **.html, .csv o .fasta**. En muchas funciones, también se puede escoger entre distintos formatos de salida .html que se pueden visualizar en el navegador. Los resultados pueden copiarse y pegarse en un archivo de Excel o Word si se quieren modificar.

Detrás de cada apartado del manual, se encuentra el **Paso a Paso** que explica gráficamente cómo llevar a cabo los análisis con cada una de las funciones.

Código de colores: en varios archivos de salida se mostrarán los residuos, porcentajes o las celdas de las tablas coloreadas según su porcentaje de frecuencia de acuerdo al siguiente código de colores:

Morado: $x = 100\%*$

Rojo: $100\% \leq x \geq 90\%$

Naranja: $90\% < x > 75\%$

Amarillo: $75\% \leq x > 50\%$

Azul: $50\% \leq x \geq 10\%$

Verde: $x < 10\%$

* el color morado corresponde al verdadero 100%, es decir, que absolutamente todas las secuencias poseen ese residuo para esa posición. Si el 100% fuera rojo, el % estaría redondeado (ej, de 99.9995 a 100%). EpiMolBio da los porcentajes con distintos números de decimales (entre 3 y 5), según la función.

El código de colores puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

| Mutaciones de Resistencia Codones MDR-IP | | |
|--|---|-----------------|
| PR_procesado_01_AE.fasta | | |
| MDR-IP Principales | | |
| Posición | Residuos | Codones Totales |
| D30N | D[GAT(0.484%)][G[GAY(0.339%)][G[GAC(1.141%)][G[RAK(0.007%)][K[KAAQ(0.004%)][N[NAT(0.004%)][N[NAT(0.004%)][G[GAK(0.004%)][G[GRT(0.022%)][T[GAW(0.004%)][T[GAT(0.004%)][G[GWT(0.004%)][Q[GGT(0.007%)] | 26845 |
| V32I | V[VGA(95.128%)][V[VGT(0.258%)][V[VTC(0.055%)][V[GTR(0.746%)][V[GTM(0.082%)][V[VTW(0.053%)][V[VGT(0.205%)][V[GTY(0.007%)][A[GCA(0.015%)][V[VTA(0.026%)][V[VTA(0.004%)][V[GTT(0.007%)][L[TTA(0.015%)][V[VTA(0.011%)][V[VGA(0.026%)][E[GAA(0.007%)][V[GYA(0.015%)][L[CTA(0.007%)][V[VTA(0.004%)][V[KTW(0.004%)][V[GKA(0.007%)] | 26849 |
| M46I | M[MATG(98.424%)][V[VTA(0.048%)][L[LTTG(0.468%)][V[VGT(0.030%)][V[VTG(0.071%)][V[VAG(0.007%)][V[VTR(0.142%)][V[VTR(0.011%)][V[RTR(0.049%)][V[VCG(0.007%)][V[VTC(0.007%)][V[VCA(0.011%)][V[VAG(0.011%)][V[VAG(0.007%)][V[VTC(0.007%)][V[VAG(0.011%)][V[VAG(0.007%)][V[VTC(0.007%)][V[VAG(0.007%)][V[VTC(0.007%)][V[VAG(0.004%)][V[VTC(0.004%)][V[VCA(0.034%)][V[VWG(0.004%)][V[VTR(0.004%)] | 26848 |
| M46L | M[MATG(98.424%)][V[VTA(0.648%)][L[LTTG(0.468%)][V[VGT(0.030%)][V[VTG(0.071%)][V[VAG(0.007%)][V[VTR(0.142%)][V[VTR(0.011%)][V[RTR(0.049%)][V[VCG(0.007%)][V[VTC(0.007%)][V[VCA(0.011%)][V[VAG(0.011%)][V[VAG(0.007%)][V[VTC(0.007%)][V[VAG(0.004%)][V[VTC(0.004%)][V[VCA(0.034%)][V[VWG(0.004%)][V[VTR(0.004%)] | 26848 |
| I47A | I[IATA(99.728%)][V[VTT(0.019%)][V[VTR(0.041%)][V[KAG(0.004%)][V[AWA(0.011%)][V[VTA(0.048%)][V[CAC(0.030%)][V[GCC(0.004%)][V[CCA(0.011%)][M[MATG(0.007%)][V[VWV(0.018%)][V[VTR(0.004%)][V[RTR(0.034%)][V[RAGG(0.004%)][V[VTA(0.004%)][V[KAA(0.011%)][F[TTT(0.004%)][V[RAGG(0.004%)][V[VTM(0.007%)][V[VTH(0.004%)][V[AKA(0.004%)][V[VTA(0.004%)] | 26848 |

Algunas funciones de EpiMolBio requieren de la introducción de una **secuencia de referencia**, ésta debe introducirse sin espacios ni saltos de línea. Generalmente los alineamientos .fasta descargados de bases de datos públicas de secuencias aparecen con saltos de línea. Una forma sencilla de eliminar los saltos de línea es copiar la secuencia de referencia del archivo del alineamiento o descargada del Genbank con herramientas tipo Pinetools (<https://pinetools.com/es/eliminar-saltos-linea>) o similar.

Las salidas tipo **.csv** pueden abrirse con Excel. Puede ocurrir que algunos caracteres especiales como las tildes no se visualicen correctamente. Para corregir esto, abrir el archivo .csv en Excel, pinchar en **Datos**, seleccionar **Obtener Datos**, seleccionar **De un archivo** y seleccionar **De texto/CSV**. Cargar el archivo .csv y transformar datos. El archivo se abrirá en el editor de Power Query. Comprobar los caracteres y seleccionar Cerrar y cargar. Si los caracteres continúan estando alterados habrá que cambiar el origen de los datos: seleccionar Configuración del origen de datos, pulsar Cambiar Origen, seleccionar opción Unicode (UTF-8), Aceptar, Cerrar y pulsar Cerrar y cargar.

Abreviaturas frecuentes:

MDR: mutaciones de resistencia

AA: aminoácido

NT: nucleótido

FUNCIONES DE EPIMOLBIO

| | |
|--|-----|
| I. VIRUS..... | 8 |
| I.1. VIH..... | 9 |
| I.1.A. MUTACIONES DE RESISTENCIA..... | 9 |
| I.1.B. OTRAS MUTACIONES POL..... | 25 |
| I.1.C. CONSERVACIÓN POL..... | 33 |
| I.2. SARS-CoV-2 RASTREADOR DE PROTEÍNAS..... | 39 |
| II. VARIABILIDAD..... | 44 |
| II.1. POLIMORFISMOS..... | 44 |
| II.2. CONSERVACIÓN..... | 81 |
| II.3. CONSENSOS..... | 92 |
| II.4. COEFICIENTE WU-KABAT..... | 100 |
| II.5. FRECUENCIA DE MUTACIÓN..... | 105 |
| III. HOMOLOGÍA..... | 110 |
| III.1. SIMILITUD..... | 110 |
| III.2. SIMILITUD PARCIAL..... | 115 |
| III.3. BÚSQUEDA DE SECUENCIAS CONSERVADAS..... | 122 |
| IV. RASTREADOR..... | 129 |
| IV.1. SIMILITUD..... | 129 |
| IV.2. FLANQUEANTES..... | 135 |
| V. ALINEAMIENTOS..... | 141 |
| V.1. ALINEAMIENTOS MÚLTIPLES..... | 141 |
| V.2. DOT PLOT..... | 146 |
| V.3. ELIMINAR INSERCIÓNES..... | 150 |
| VI. HERRAMIENTAS..... | 154 |
| VI.1. EDICIÓN DE ARCHIVOS..... | 154 |
| VI.2. FILTROS..... | 172 |
| VI.3. TRADUCCIÓN..... | 187 |
| VI.4. CONTAR SECUENCIAS..... | 192 |
| VI.5. PROGRAMAR FUNCIONES..... | 198 |

Resumen de las funciones de EpiMolBio

| Sección | Sub-sección | Función |
|-------------------|--|--|
| VIRUS | VIH | |
| | Mutaciones de Resistencia (Adquiridas* y Transmitidas) | Calcula el porcentaje de MDR adquiridas/transmitidas, respecto a una secuencia de referencia, a partir de secuencias de AA de las proteínas de Pol del VIH y la Cápside. |
| | Otras Mutaciones Pol | Detecta cualquier mutación (no sólo MDR) del VIH-1 o VIH-2 a partir de secuencias de las proteínas de Pol, reportando su porcentaje con respecto a la secuencia de referencia. |
| | Conservación Pol | Genera una tabla con el AA más prevalente y su porcentaje, permitiendo conocer el residuo más conservado para cada posición de la proteína Pol seleccionada. |
| VARIABILIDAD | SARS-CoV-2 Rastreador de Proteínas | Genera secuencias en formato .fasta de las proteínas que se escogen del SARS-CoV-2 a partir de genomas completos, en NT o AA. |
| | Polimorfismos* | Posiciones Mutadas y Tabla de Mutaciones permiten la detección de polimorfismos informando de su localización y frecuencia de aparición utilizando como referencia cualquier secuencia introducida por el usuario. Marcadores permite detectar las mutaciones exclusivas de cada archivo en comparación al resto de archivos introducidos como entrada. Mutaciones Múltiples permite detectar y conocer la frecuencia de aparición de combinaciones de mutaciones. Mutaciones por Posición permite detectar y conocer la frecuencia de aparición de residuos en una posición o varias posiciones combinadas. |
| | Conservación* | Determina el grado de conservación de secuencias de interés informando del residuo más prevalente y su porcentaje, generando secuencias consenso. La función "Codones" presenta el codón más conservado de una secuencia de NT. |
| | Consensos | Genera secuencias consenso y consensos de consenso realizando varias rondas de análisis. |
| | Coeficiente de Wu-Kabat | Genera el coeficiente de variabilidad de Wu-Kabat de secuencias de proteínas para estudiar la susceptibilidad de una posición de AA a los reemplazos evolutivos. |
| HOMOLOGÍA | Frecuencia de Mutación | Genera una serie de parámetros relacionados con la frecuencia de aparición de mutaciones en un grupo de secuencias como la frecuencia de mutación, el porcentaje de conservación y las mutaciones medias por secuencia. |
| | Similitud | Busca una secuencia problema introducida por el usuario entre las del archivo de entrada obteniendo la proporción de secuencias por archivo que contienen la secuencia problema. |
| | Similitud Parcial | Compara una secuencia introducida por el usuario con las secuencias de entrada para buscar regiones similares entre ambas, definiendo el porcentaje de similitud. |
| RASTREADOR | Búsqueda de Secuencias Conservadas | Busca fragmentos de secuencias conservadas a partir de un conjunto de secuencias de entrada pudiendo buscar en una región concreta, escoger la longitud del fragmento, y el porcentaje de conservación. |
| | Similitud | Busca secuencias de interés dentro de un conjunto de secuencias más largas a partir de una secuencia de referencia. |
| | Flanqueantes | Busca proteínas dentro de un conjunto de secuencias genómicas completas utilizando las secuencias flanqueantes de la proteína buscada. |
| ALINEAMIENTOS | Alineamientos Múltiples | Alinea secuencias de AA y NT utilizando el programa MUSCLE v3.8.31. |
| | Dot Plot | Genera una representación gráfica donde se comparan secuencias trazando puntos en una matriz bidimensional, con cada eje representando una secuencia. |
| | Eliminar Inserciones | Elimina automáticamente las inserciones de una secuencia con respecto a una referencia con gaps tras haber realizado el alineamiento. |
| HERRAMIENTAS | Edición de Archivos | |
| | Fusionar Archivos | Combina múltiples archivos .fasta en uno solo. |
| | Secuencias Únicas | Elimina las secuencias duplicadas de uno o varios archivos .fasta. |
| | Búsqueda de Secuencias | Filtrá secuencias de archivos .fasta que contienen una o varias mutaciones seleccionadas por el usuario. |
| | Buscar y Reemplazar | Reemplaza una serie de caracteres por otros en el encabezado y/o la secuencia genética de uno o varios archivos .fasta. |
| | Filtros | |
| | Filtrado por Encabezado | Filtrá uno o varios archivos en formato ".fasta" utilizando los parámetros de sus encabezados. |
| | Filtro Específico | Filtrá secuencias de archivos en formato ".fasta" que tienen un conjunto específico de caracteres en sus encabezados. |
| | Filtro de Secuencias Parciales | Filtrá secuencias de archivos en formato ".fasta" en función de su calidad, según la cantidad de "?" (secuencias en AA) o "N" (secuencias en NT) que contienen. |
| | Traducción | Traduce secuencias .fasta de NT a AA. |
| Contar Secuencias | | Cuenta el número total de secuencias en uno o varios archivos .fasta o cuántas de esas secuencias contienen mutaciones. |
| | Programación de Funciones | Automatiza las funciones del programa encadenándolas para que se ejecuten secuencialmente sin intervención manual. |

* Análisis individual y de codones. Abreviaturas: MDR, mutaciones de resistencia;AA, aminoácidos; NT, nucleótidos.

I. VIRUS

En esta sección se encuentran las herramientas específicas para el análisis del Virus de la Inmunodeficiencia Humana (VIH) y del Coronavirus del Síndrome Respiratorio Agudo Grave de Tipo 2 (SARS-CoV-2).

I.1. VIH

I.1.A) MUTACIONES DE RESISTENCIA

Esta función permite la detección de mutaciones de resistencia (MDR) del VIH a partir de secuencias de las proteínas de Pol y la Cápside en aminoácidos, obteniendo el porcentaje de MDR y su clasificación con respecto a la secuencia de referencia de VIH-1 o de VIH-2. La secuencia de referencia para VIH-1 que emplea EpiMolBio es HXB2 (NCBI K03455.1) y para VIH-2 es ALI (NCBI AF082339). EpiMolBio no detecta las delecciones o inserciones que afecten a la susceptibilidad a fármacos antirretrovirales.

El análisis puede realizarse por **mutaciones individuales**, sólo en aminoácidos, o por **codones** o tripletes en nucleótidos.

Mutaciones individuales:

En el análisis de **mutaciones individuales** se pueden estudiar tanto las **MDR adquiridas** contempladas en Stanford HIV Drug Resistance Database v9.7 **para VIH-1** y las **MDR adquiridas** contempladas en HIV-2 EU Tool v.2 2015, Charpentier et al. 2015, Tzou et al. 2020, Troyano-Hernández P et al. 2021 **para VIH-2**. También se puede analizar la presencia de las **MDR transmitidas para VIH-1 o SDRM** según el listado de la OMS, Bennet et al. 2009 y según el listado de Stanford HIV Drug Resistance Database v9.7 (<https://cms.hivdb.org/prod/downloads/resistance-mutation-handout/resistance-mutation-handout.pdf>), con última actualización el 9-11-2024. La lista completa de mutaciones integradas en el programa EpiMolBio se puede consultar en el **Anexo I**.

Se analizan las mutaciones detectadas en las proteínas del gen *pol* del VIH: Proteasa, Retrotranscriptasa o Integrasa, y en la Cápside. EpiMolBio excluye de manera automática tanto los gaps (-) como las interrogaciones (?) del análisis. Los codones de stop se indican con un asterisco negro (*).

En el caso de **MDR adquiridas** habrá que seleccionar el tipo de VIH que se va a analizar para establecer la secuencia de referencia: **VIH-1** o **VIH-2**.

En ambos casos, el **archivo de entrada** debe ser la **carpeta** que contenga exclusivamente las secuencias alineadas de la proteína de Pol que se quiera analizar (Proteasa, Retrotranscriptasa, Integrasa o Cápside) en aminoácidos y en formato **.fasta**. Esta carpeta puede contener un solo archivo o varios archivos .fasta si queremos analizar lo mismo en distintos grupos de secuencias (ej.: archivos divididos por variantes del VIH, país, año, etc).

El **archivo de salida** será un archivo con extensión **.html**. Habrá que seleccionar la carpeta de salida donde se quiere que aparezcan los archivos generados y nombrar los archivos escribiendo .html al final.

En el campo “**Tipo de MDR**” para **Mutaciones Adquiridas** o “**Tipo de SDRM**” en **Mutaciones Transmitidas**, habrá que seleccionar el **tipo de mutación** que se desea estudiar, seleccionando entre mutaciones frente a:

IP: inhibidores de la Proteasa (entrada: Proteasa)

ITIAN: inhibidores de la Transcriptasa Inversa análogos de los nucleós/tidos (entrada: Transcriptasa Inversa)

ITINAN: inhibidores de la Transcriptasa Inversa no análogos de los nucleós/tidos (entrada: Transcriptasa Inversa)

INI: inhibidores de la Integrasa (entrada: Integrasa)

ICA: inhibidores de la Cápside (entrada: Cápside)

En el campo “**Formato de salida**”, escoger entre los tres tipos de formato de salida: **lista**, **tabla** y **tabla resumen**.

En los tres casos se obtendrán las **mutaciones** detectadas en las secuencias introducidas en el archivo de entrada, su **clasificación** según la clasificación de Stanford v9.7 (<https://hivdb.stanford.edu/page/release-notes/#drm.classification>) y su **porcentaje** con respecto al total de posiciones válidas según el **código de colores** descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

1.- Lista:

En este formato se muestran solo las posiciones que contienen mutaciones de resistencia. Las posiciones se describen bajo la columna “Posición”. En la columna “Residuos” se pueden ver todos los residuos encontrados en las secuencias analizadas y su porcentaje coloreado según el código de colores. La MDR o SDRM estará indicada con un asterisco de color rojo (*). Al final de cada fila, en la columna “Posiciones Totales”, aparece el número total de secuencias válidas para esa posición que están presentes en el archivo analizado.

En la parte superior aparecerá indicado el título del análisis, el nombre del archivo de entrada y la clasificación de las MDR.

Ejemplo de formato de salida Lista para el análisis de mutaciones de resistencia adquiridas:

| Lista Mutaciones de Resistencia Adquiridas MDR-IP VIH-1 | | |
|---|--|--------------------|
| PR_procesado_traducido_01_AE.fasta | | |
| MDR-IP PRINCIPALES | | |
| Posición | Residuos | Posiciones Totales |
| D30 | D(99.985%) K(0.004%) N(0.004%*) G(0.007%) | 26741 |
| V32 | V(99.929%) A(0.015%) I(0.026%*) L(0.023%) E(0.008%) | 26584 |
| M46 | M(98.733%) I(0.661%*) L(0.531%*) V(0.064%) R(0.011%) | 26764 |
| I47 | I(99.903%) K(0.015%) V(0.048%*) A(0.015%*) M(0.007%) R(0.007%) F(0.004%) | 26814 |
| G48 | G(99.828%) S(0.007%*) R(0.037%) V(0.101%*) M(0.015%*) I(0.004%) Q(0.004%*) E(0.004%) | 26707 |

Ejemplo de formato de salida Lista para el análisis de mutaciones de resistencia transmitidas:

| Lista Mutaciones de Resistencia Transmitidas SDRM-IP OMS | | |
|--|---|--------------------|
| PR_procesado_traducido_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| L23 | L(99.935%) I(0.019%*) F(0.015%) V(0.019%) Q(0.004%) P(0.004%) S(0.004%) | 26251 |
| L24 | L(99.940%) I(0.015%*) V(0.019%) S(0.019%) F(0.007%) | 26714 |
| D30 | D(99.985%) K(0.004%) N(0.004%*) G(0.007%) | 26741 |
| V32 | V(99.929%) A(0.015%) I(0.026%*) L(0.023%) E(0.008%) | 26584 |
| M46 | M(98.733%) I(0.661%*) L(0.531%*) V(0.064%) R(0.011%) | 26764 |

2.- Tabla:

En este formato se muestra, en la primera fila, el título del análisis. En caso de analizar MDR, debajo aparece el tipo de MDR según su clasificación. Tanto en MDR como en SDRM, en la primera columna aparece el nombre de los archivos de entrada empleados para generar la tabla. En el resto de columnas se muestra la MDR o SDRM detectada con la celda coloreada según el código de colores, así como su porcentaje de aparición en cada archivo de entrada.

Ejemplo de formato de salida Tabla para el análisis de mutaciones de resistencia adquiridas:

| | Principales | | | | | | | | | | | | | | | | | | |
|-------------------------------------|-------------|--------|--------|---------|--------|--------|--------|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | D30N | V32I | M46I | M46L | I47A | I47V | G48A | G48L | G48M | G48Q | G48S | G48T | G48V | I50L | I50V | I54A | I54L | I54M | I54S |
| PR_procesado_traducido_01_AE.fasta | 0.004% | 0.026% | 0.661% | 0.531% | 0.015% | 0.048% | | | 0.015% | 0.004% | 0.007% | | 0.101% | | 0.028% | 0.011% | 0.023% | 0.011% | 0.015% |
| PR_procesado_traducido_02_AG.fasta | 0.063% | 0.021% | 0.859% | 0.178% | | 0.084% | 0.032% | | 0.042% | | | 0.011% | 0.011% | 0.010% | | | 0.053% | 0.021% | |
| PR_procesado_traducido_03_A6B.fasta | | 0.990% | 1.303% | 0.651% | 0.649% | | | | | | | | | 0.324% | 0.647% | | | | |
| PR_procesado_traducido_04_cpx.fasta | 6.667% | | | 13.333% | | | | | | | | | 6.667% | | | | | | 7.143% |
| PR_procesado_traducido_05_DF.fasta | | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_06_cpx.fasta | | | 1.754% | 0.270% | | 0.268% | | | | | | | 0.135% | | | | | | |
| PR_procesado_traducido_07_BC.fasta | 0.018% | | 0.055% | 0.037% | | | | | | | | | | 0.009% | | | | | |
| PR_procesado_traducido_08_BC.fasta | 0.043% | | 0.128% | | 0.043% | | | | | | | | 0.128% | | | | | | |

Ejemplo de formato de salida Tabla para el análisis de mutaciones de resistencia transmitidas:

| | Tabla Mutaciones de Resistencia Transmitidas SDRM-IP OMS | | | | | | | | | | | | | | | | | | | | |
|-------------------------------------|--|--------|--------|--------|---------|--------|--------|--------|--------|--------|---------|--------|---------|--------|--------|--------|--------|--------|--------|--------|--------|
| | L23I | L24I | D30N | V32I | M46I | M46L | I47V | I47A | G48V | G48M | I50V | I50L | F53L | F53Y | I54V | I54L | I54M | I54A | I54T | I54S | |
| PR_procesado_traducido_01_AE.fasta | 0.019% | 0.015% | 0.004% | 0.028% | 0.661% | 0.531% | 0.048% | 0.015% | 0.161% | 0.015% | 0.028% | | 0.119% | 0.022% | 0.222% | 0.023% | 0.011% | 0.011% | 0.015% | 0.037% | 0.064% |
| PR_procesado_traducido_02_AG.fasta | 0.021% | 0.021% | 0.063% | 0.021% | 0.859% | 0.178% | 0.084% | | 0.011% | 0.042% | | 0.010% | 0.115% | 0.115% | 0.465% | 0.053% | 0.021% | | | 0.159% | |
| PR_procesado_traducido_03_A6B.fasta | | | | 0.990% | 1.303% | 0.651% | | 0.649% | | | 0.547% | 0.524% | 0.378% | | 0.329% | | | | | 0.324% | |
| PR_procesado_traducido_04_cpx.fasta | 6.667% | 6.667% | | | 13.333% | | | 0.067% | | | 13.333% | | 26.571% | | | | | | 7.143% | | |
| PR_procesado_traducido_05_DF.fasta | | | | | | | | | | | | | | | | | | | | 0.045% | |
| PR_procesado_traducido_06_cpx.fasta | 0.137% | | | | 1.754% | 0.270% | 0.268% | | | | 0.135% | 0.289% | 0.269% | 1.207% | | | | | | | |
| PR_procesado_traducido_07_BC.fasta | | 0.018% | | | 0.055% | 0.037% | | | | 0.009% | | | | | | | | | | 0.045% | |
| PR_procesado_traducido_08_BC.fasta | 0.043% | 0.043% | | 0.128% | | 0.043% | | | 0.128% | | | 0.043% | | 1.064% | 2.151% | | | | | | |
| PR_procesado_traducido_09_cpx.fasta | | | | | | | | | | | | | | | | | | | | | |

3.- Tabla Resumen:

En esta tabla se muestra, en la primera fila, el título del análisis. Debajo, en la primera columna “Archivo”, aparece el nombre de los archivos de entrada empleados para generar el análisis. En las siguientes columnas, se muestra el tipo de SDRM o MDR según su clasificación, con los residuos encontrados correspondientes a las MDR o SDRM coloreados según su porcentaje siguiendo el código de colores anteriormente descrito.

Ejemplo de formato de salida Tabla Resumen para el análisis de mutaciones de resistencia adquiridas:

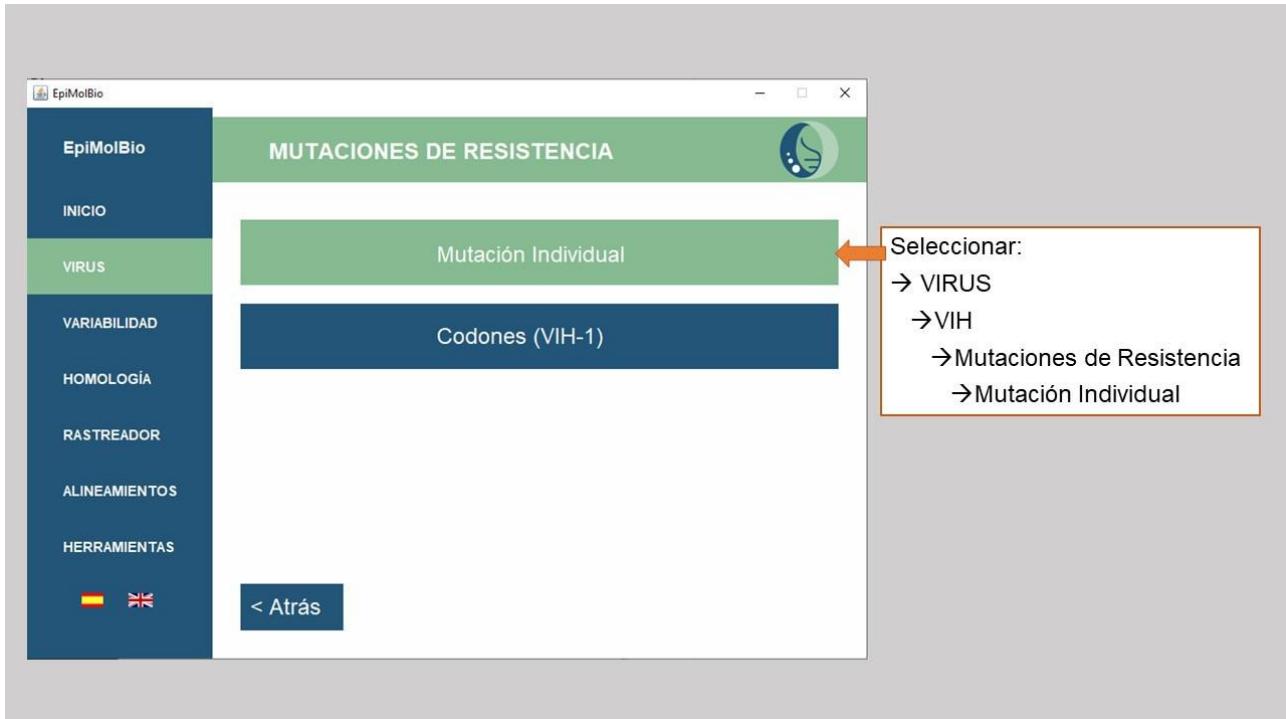
| Tabla Resumen Mutaciones de Resistencia Adquiridas MDR-IP VIH-1 | | | |
|---|--|---|---|
| Archivo | MDR IP | | |
| | Principales | Accesorias | Otras |
| PR_procesado_traducido_01_AE.fasta | D30N, V32I, M46IL, I47AV, G48MQSV, I50V, I54ALMSV, L76V, V82AFLMST, I84V, N88GS, L90M | L10F, K20T, L23I, L24FI, L33F, K43T, M46V, F53LY, Q58E, G73DSTV, T74P, N83D, N88D, L89TV | L10IRVY, V11IL, K20IMRV, L33IV, A71ITV, T74S, V82I, I85V, L89IM |
| PR_procesado_traducido_02_AG.fasta | D30N, V32I, M46IL, I47V, G48AMTV, I50L, I54LMV, L76V, V82ACFLST, I84ACV, N88ST, L90M | L10F, K20T, L23I, L24I, L33F, K43T, M46V, F53LY, Q58E, G73ADSV, T74P, N83D, N88D, L89TV | L10IRVY, V11IL, K20IMRV, L33IV, A71ITV, T74S, V82I, I85V, L89IM |

Ejemplo de formato de salida Tabla Resumen para el análisis de mutaciones de resistencia transmitidas:

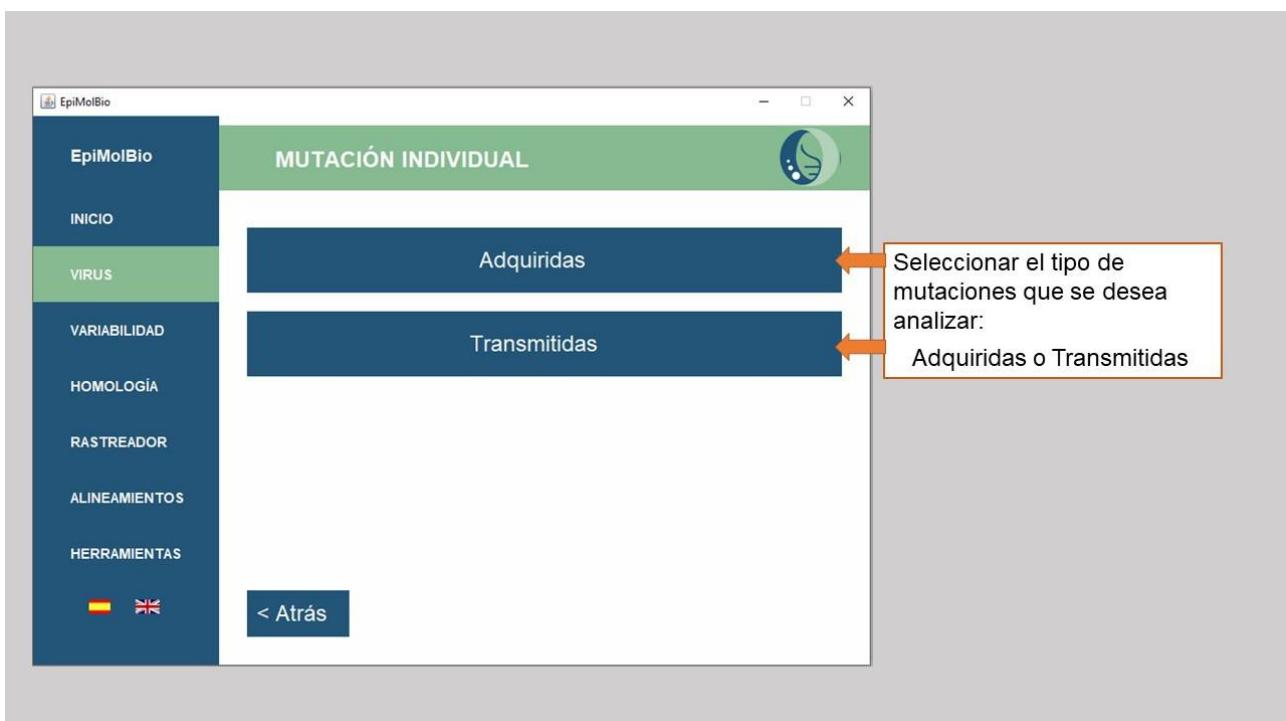
| Tabla Resumen Mutaciones de Resistencia Transmitidas SDRM-IP OMS | |
|--|---|
| Archivo | SDRM IP OMS |
| PR_procesado_traducido_01_AE.fasta | L23I, L24I, D30N, V32I, M46IL, I47VA, G48VM, I50V, F53LY, I54VLMAS, G73ST, L76V, V82ATFSML, N83D, I84V, I85V, N88DS, L90M |
| PR_procesado_traducido_02_AG.fasta | L23I, L24I, D30N, V32I, M46IL, I47V, G48VM, I50L, F53LY, I54VLM, G73SA, L76V, V82ATFSCL, N83D, I84VAC, I85V, N88DS, L90M |
| PR_procesado_traducido_03_A6B.fasta | V32I, M46IL, I47A, I50VL, F53L, I54V, G73S, L76V, V82AT, I84V, L90M |

Paso a paso:

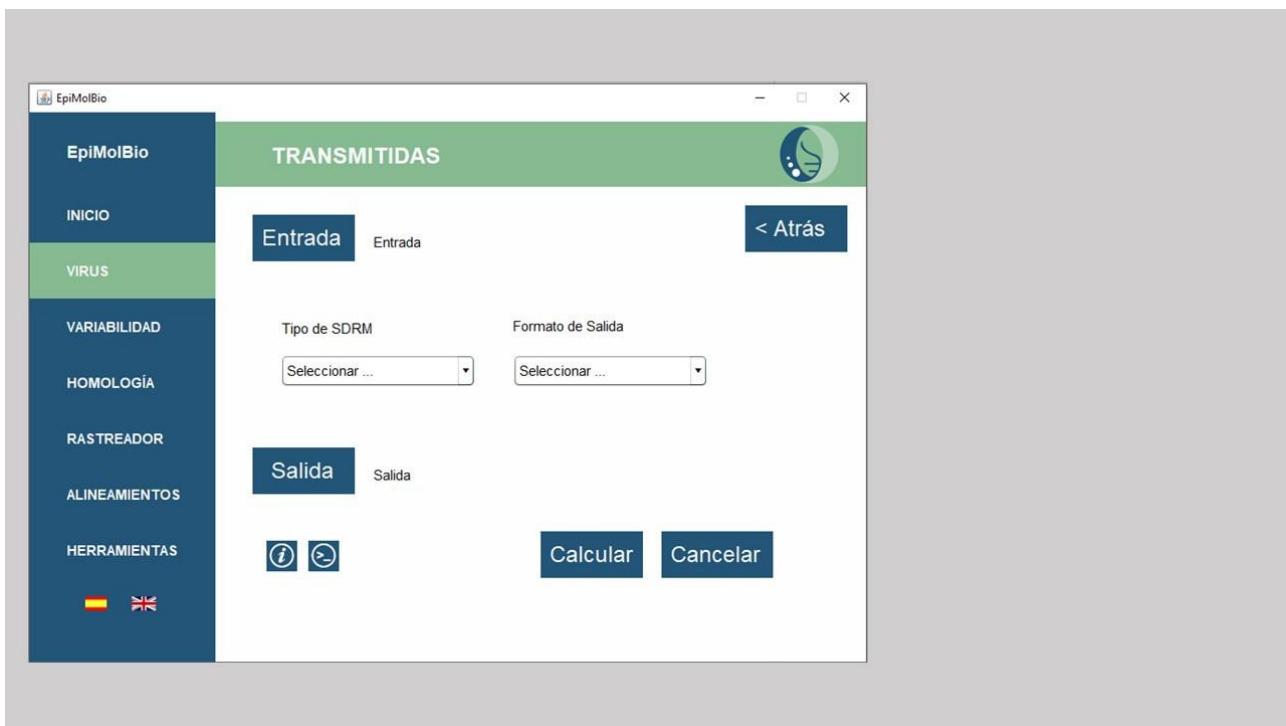
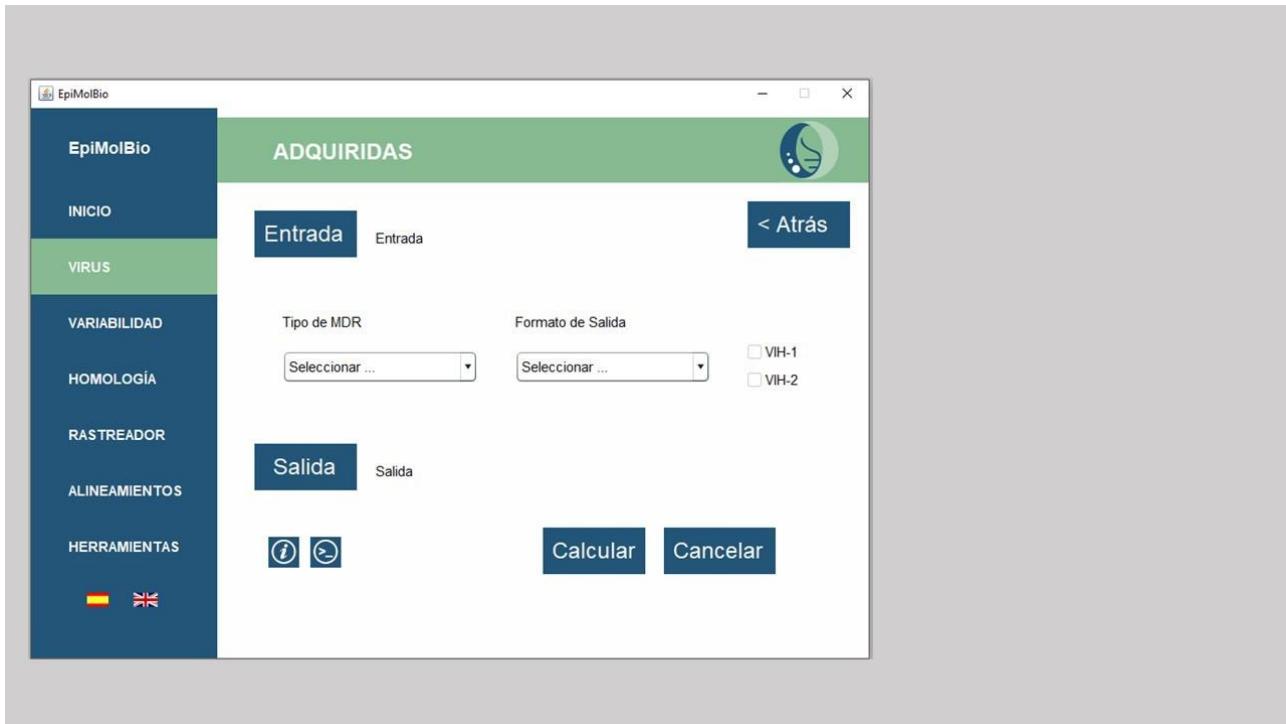
1)



2)

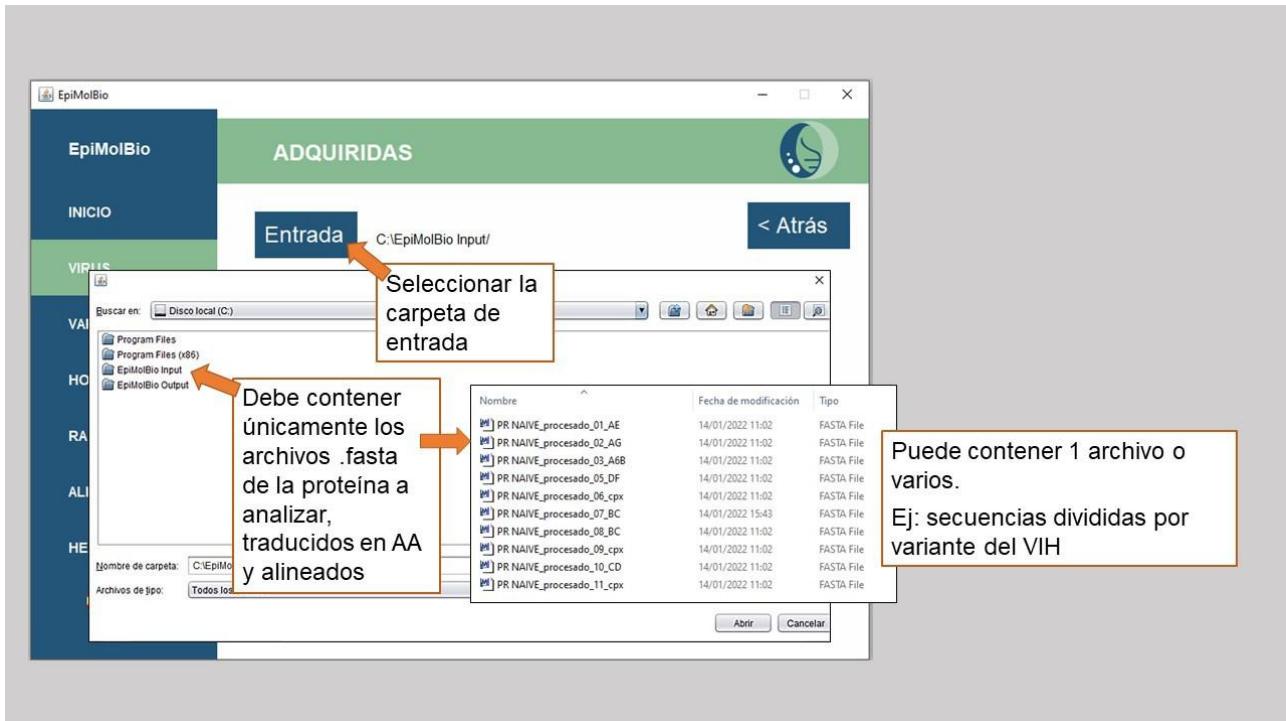


3)

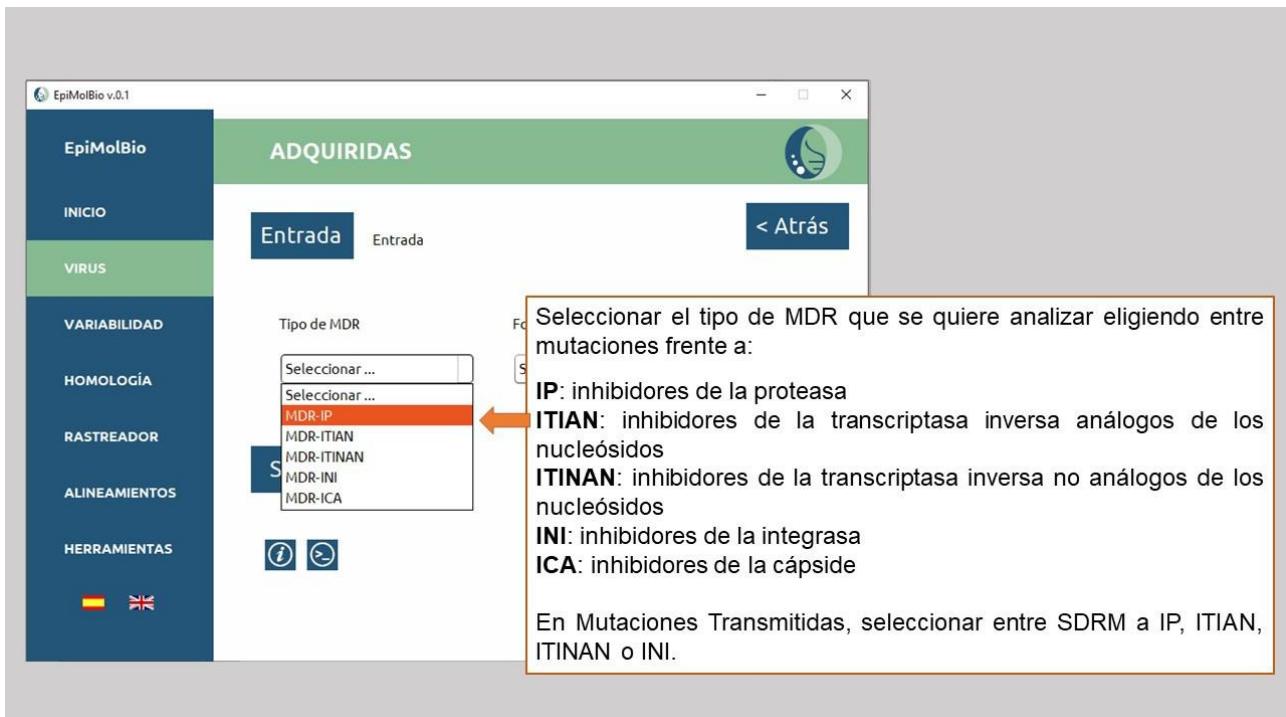


Los siguientes pasos son idénticos para ambos tipos de mutaciones excepto el paso 7, exclusivo para la detección de MDR adquiridas.

4)



5)



6)

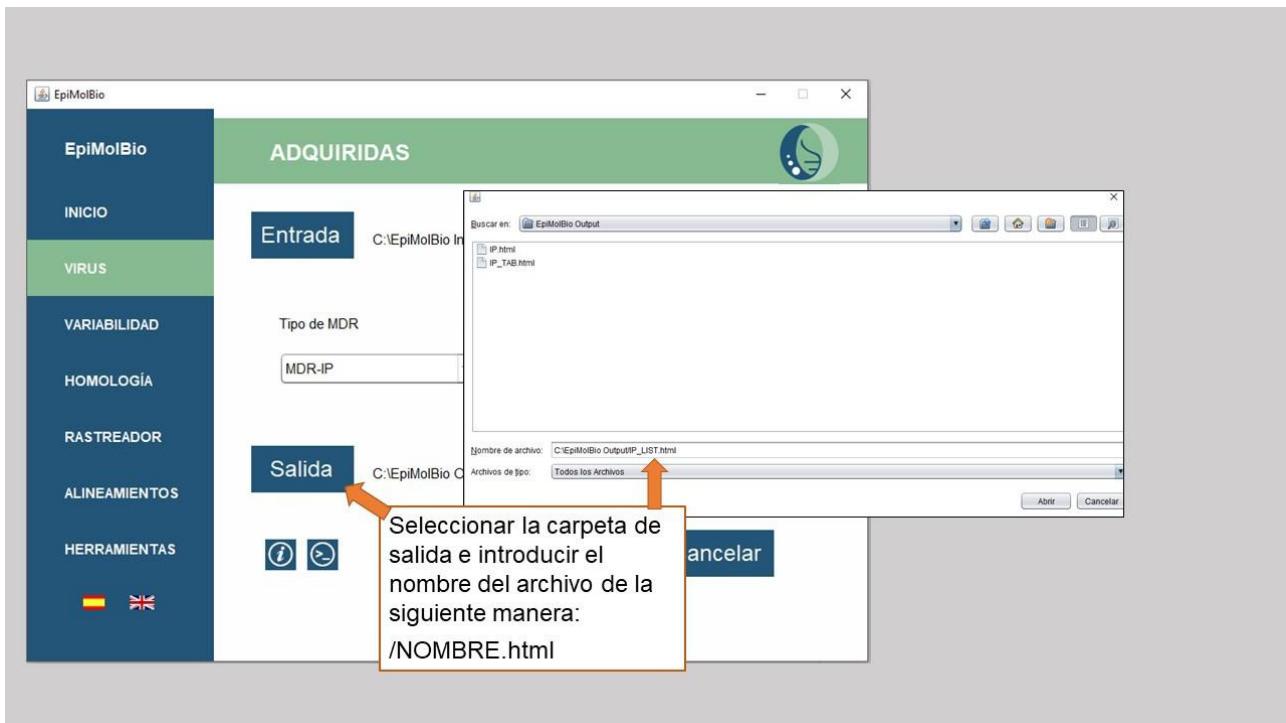


7)

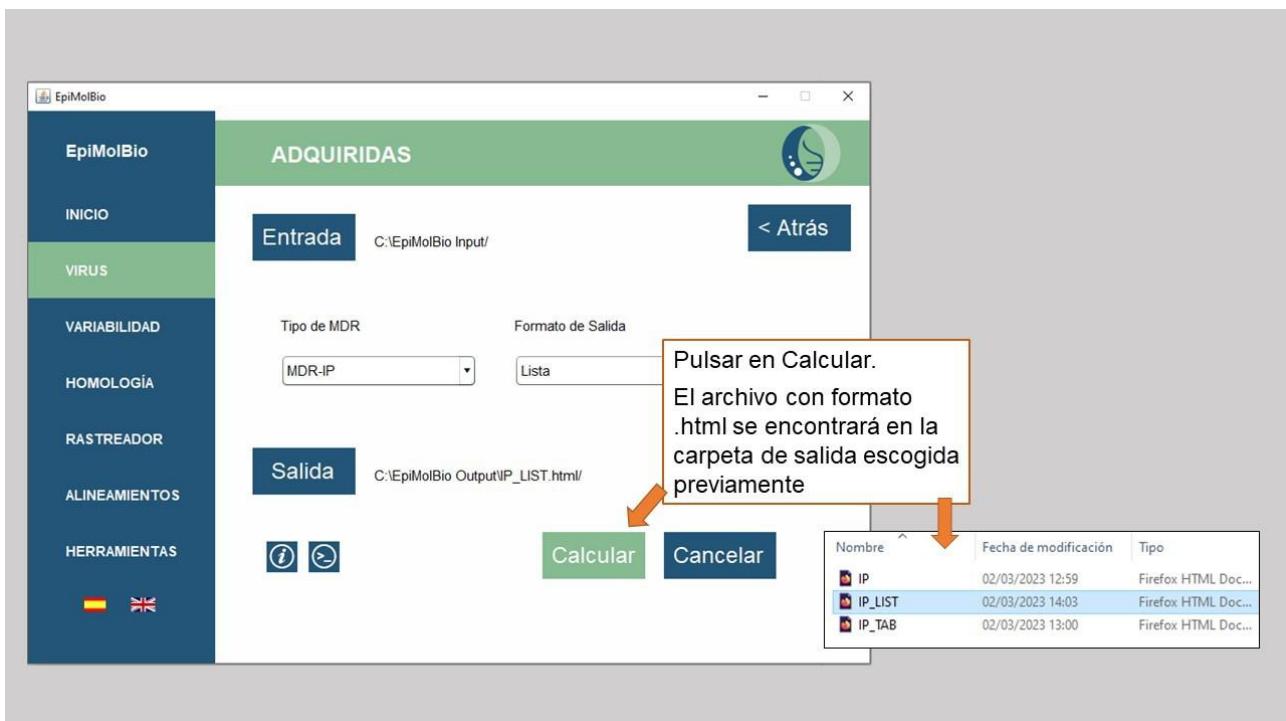


Este paso se omite en el análisis de resistencias transmitidas, ya que la secuencia de referencia es la del VIH-1 por defecto.

8)



9)



Mutaciones codones:

En el análisis de “**Mutaciones de Resistencia**” opción “codones” se pueden estudiar los codones que, al traducirse, generan MDR del VIH-1 contempladas en la Stanford HIV Drug Resistance Database v9.7, cuantificando la frecuencia de aparición de cada codón y aquellos que generan MDR. Se analizan las mutaciones detectadas en las proteínas del gen *pol* del VIH-1: Proteasa, Retrotranscriptasa o Integrasa y de la Cápside. Tanto los gaps (-) como las “N” son excluidas del análisis. La secuencia de referencia para VIH-1 que emplea EpiMolBio es HXB2 (NCBI K03455.1).

El **archivo de entrada** debe ser la **carpeta** que contenga exclusivamente las secuencias alineadas de la proteína de Pol o la Cápside que se quiera analizar en nucleótidos y en formato .fasta. Esta carpeta puede contener un solo archivo o varios archivos .fasta si se quiere analizar lo mismo en distintos grupos de secuencias (ej.: archivos divididos por variantes del VIH, país de origen, año, etc.).

En el campo “**Tipo de MDR**” habrá que seleccionar el **tipo de mutación** que se desea estudiar, seleccionando entre mutaciones frente a:

IP: inhibidores de la Proteasa (entrada: Proteasa)

ITIAN: inhibidores de la Transcriptasa Inversa análogos de los nucleós/tidos (entrada: Transcriptasa Inversa)

ITINAN: inhibidores de la Transcriptasa Inversa no análogos de los nucleós/tidos (entrada: Transcriptasa Inversa)

INI: inhibidores de la Integrasa (entrada: Integrasa)

ICA: inhibidores de la Cápside (entrada: Cápside)

Se podrá realizar el estudio de la secuencia completa o buscar mutaciones específicas en uno o varios codones seleccionando “**Seleccionar Mutación**” y añadiendo una o varias MDR válidas en el campo “**Mutaciones**”. En caso de introducir más de una, deben estar separadas por “,” sin espacios y en aminoácidos (ej.: V32I,I50L). Si no se selecciona ninguna mutación concreta, se mostrarán todas las posiciones que contengan mutaciones de resistencia.

En ambos casos se obtendrán las **mutaciones** detectadas en las secuencias introducidas en el archivo de entrada, su **clasificación** según la clasificación de Stanford v9.7 y su **porcentaje** con respecto al total de secuencias analizadas según el **código de colores** descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

El **archivo de salida** será un archivo con extensión **.html**. Habrá que seleccionar la carpeta de salida donde se quiere que aparezcan los archivos generados y nombrar los archivos escribiendo .html al final, excepto si se selecciona mutación, en cuyo caso el archivo se nombra automáticamente según las MDR introducidas en “**Mutaciones**”. Por cada MDR introducida se generará un archivo con extensión .html.

En el archivo de salida aparece, en la parte superior, el título del análisis, seguido del nombre del archivo de entrada y el tipo de MDR analizada (este último no aparece si se selecciona mutación). En la columna “Posición” aparecen las posiciones que contienen mutaciones de resistencia o las posiciones de las mutaciones introducidas en el campo “Mutaciones” si se selecciona mutación. En la columna “Residuos” aparecen todos los

residuos encontrados para esa posición junto al codón que lo codifica y el porcentaje de aparición de dicho codón coloreado según el código de colores. Las MDR se indican con un asterisco rojo (*). Los codones no codificantes estarán indicados con una interrogación (?). En la columna “Codones Totales” se describe el número total de secuencias válidas para esa posición que están presentes en el archivo analizado. Si no se detecta ninguna mutación en alguna de las filas, el archivo de salida muestra lo que hay en esa posición aunque no este mutado.

Ejemplo de formato de salida para el análisis de mutaciones de resistencia codones sin seleccionar mutación:

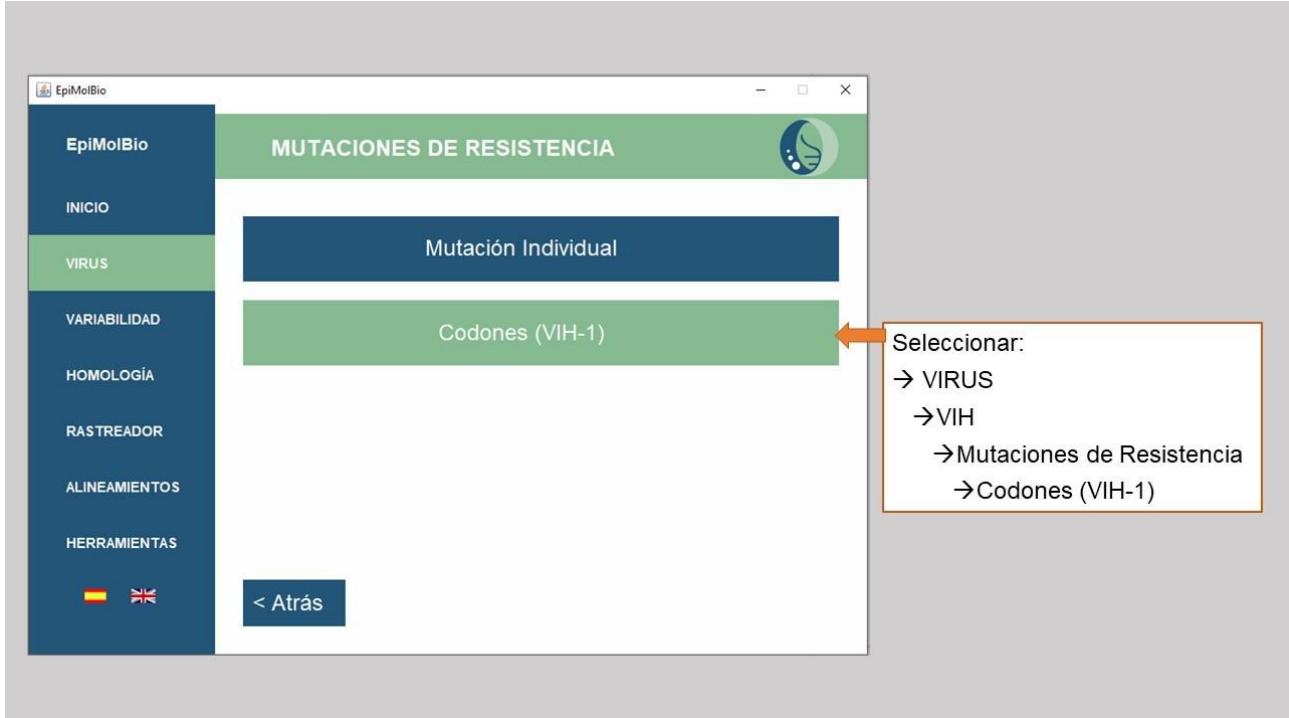
| Mutaciones de Resistencia Codones MDR-IP | | |
|--|---|-----------------|
| PR_procesado_01_AE.fasta | | |
| MDR-IP Principales | | |
| Posición | Residuos | Codones Totales |
| D30N | D[GAT(98.484%) ?[GAY(0.339%)] D[GAC(1.114%)] ?[RAK(0.007%)] K[AAG(0.004%)] N[AAT [*] (0.004%)] ?[RAT(0.004%)] ?[GAK(0.004%)] ?[GRT(0.022%)] ?[GAW(0.004%)] ?[GMT(0.004%)] ?[GWT(0.004%)] G[GGT(0.007%)] | 26845 |
| V32I | V[GTA(95.128%)] V[GTG(3.356%)] V[GTC(0.253%)] ?[GTR(0.756%)] ?[GTM(0.082%)] ?[GTVW(0.063%)] V[GTT(0.205%)] ?[GTY(0.007%)] A[GCA(0.015%)] I[ATA [*] (0.026%)] ?[RTA(0.004%)] ?[GTD(0.007%)] L[TTA(0.015%)] ?[KTA(0.011%)] ?[GWA(0.026%)] E[GAA(0.007%)] ?[GYA(0.015%)] L[CTA(0.007%)] ?[STA(0.004%)] ?[KTW(0.004%)] ?[GKA(0.007%)] | 26849 |
| M46I | M[ATG(98.424%)] I[ATA [*] (0.648%)] L[TTG(0.466%)] V[GTG(0.030%)] ?[WTG(0.071%)] ?[AYG(0.007%)] ?[ATR(0.142%)] ?[TTR(0.011%)] ?[RTG(0.045%)] L[CTG(0.030%)] I[ATT [*] (0.011%)] L[TTA(0.030%)] ?[AKG(0.011%)] ?[MTG(0.007%)] ?[TYG(0.004%)] L[CTA(0.004%)] R[AGG(0.007%)] R[AGA(0.004%)] ?[AWG(0.004%)] ?[RTA(0.004%)] V[GTA(0.034%)] ?[WWG(0.004%)] ?[RTR(0.004%)] | 26848 |

Ejemplo de formato de salida para el análisis de mutaciones de resistencia codones seleccionando la mutación M46I:

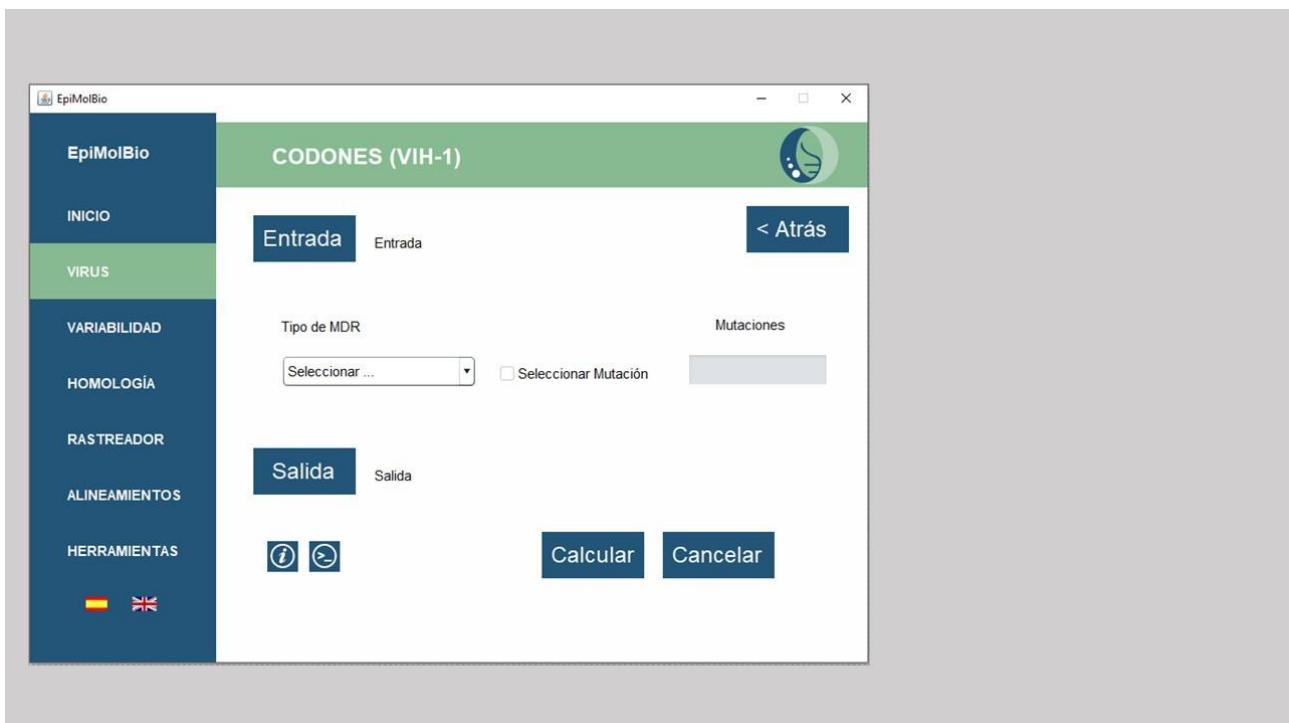
| Mutaciones de Resistencia Codones MDR-IP | | |
|--|---|-----------------|
| PR_procesado_01_AE.fasta | | |
| Posición | Residuos | Codones Totales |
| M46I | M[ATG(98.424%)] I[ATA [*] (0.648%)] L[TTG(0.466%)] V[GTG(0.030%)] ?[WTG(0.071%)] ?[AYG(0.007%)] ?[ATR(0.142%)] ?[TTR(0.011%)] ?[RTG(0.045%)] L[CTG(0.030%)] I[ATT [*] (0.011%)] L[TTA(0.030%)] ?[AKG(0.011%)] ?[MTG(0.007%)] ?[TYG(0.004%)] L[CTA(0.004%)] R[AGG(0.007%)] R[AGA(0.004%)] ?[AWG(0.004%)] ?[RTA(0.004%)] V[GTA(0.034%)] ?[WWG(0.004%)] ?[RTR(0.004%)] | 26848 |
| PR_procesado_02_AG.fasta | | |
| Posición | Residuos | Codones Totales |
| M46I | M[ATG(98.528%)] I[ATA [*] (0.835%)] L[TTG(0.157%)] V[GTG(0.063%)] ?[ATR(0.198%)] ?[WTG(0.042%)] I[ATT [*] (0.010%)] ?[TTR(0.021%)] V[GTA(0.021%)] I[ATC [*] (0.010%)] ?[AKG(0.010%)] ?[MTG(0.010%)] ?[AYG(0.010%)] ?[RTG(0.010%)] ?[RTA(0.010%)] L[TTA(0.010%)] ?[AWG(0.010%)] L[CTG(0.010%)] K[AAG(0.021%)] ?[WTR(0.010%)] | 9577 |

Paso a paso:

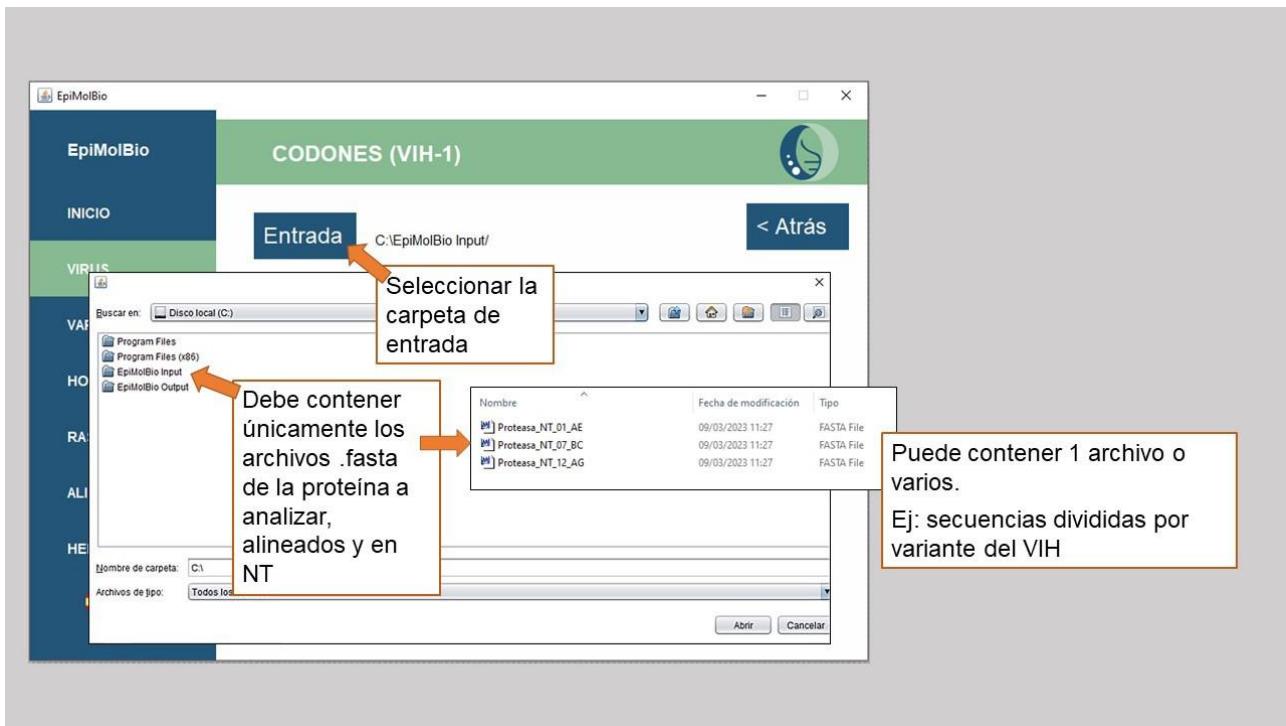
1)



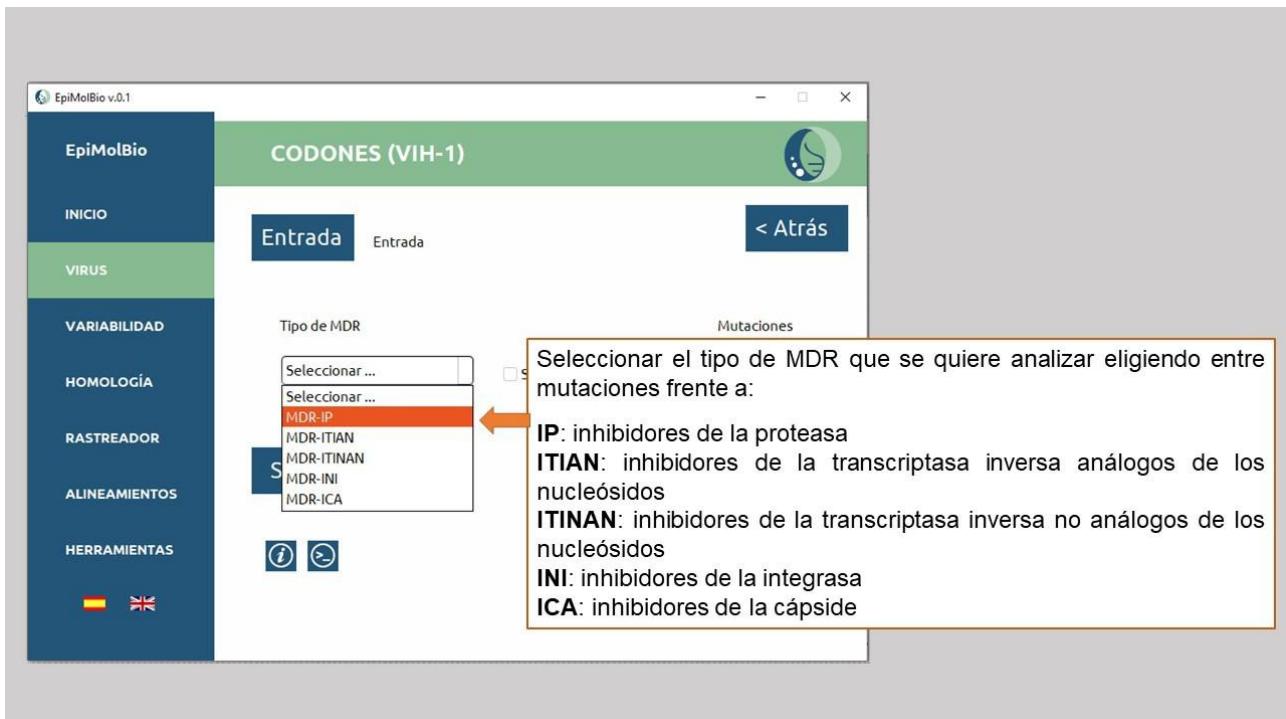
2)



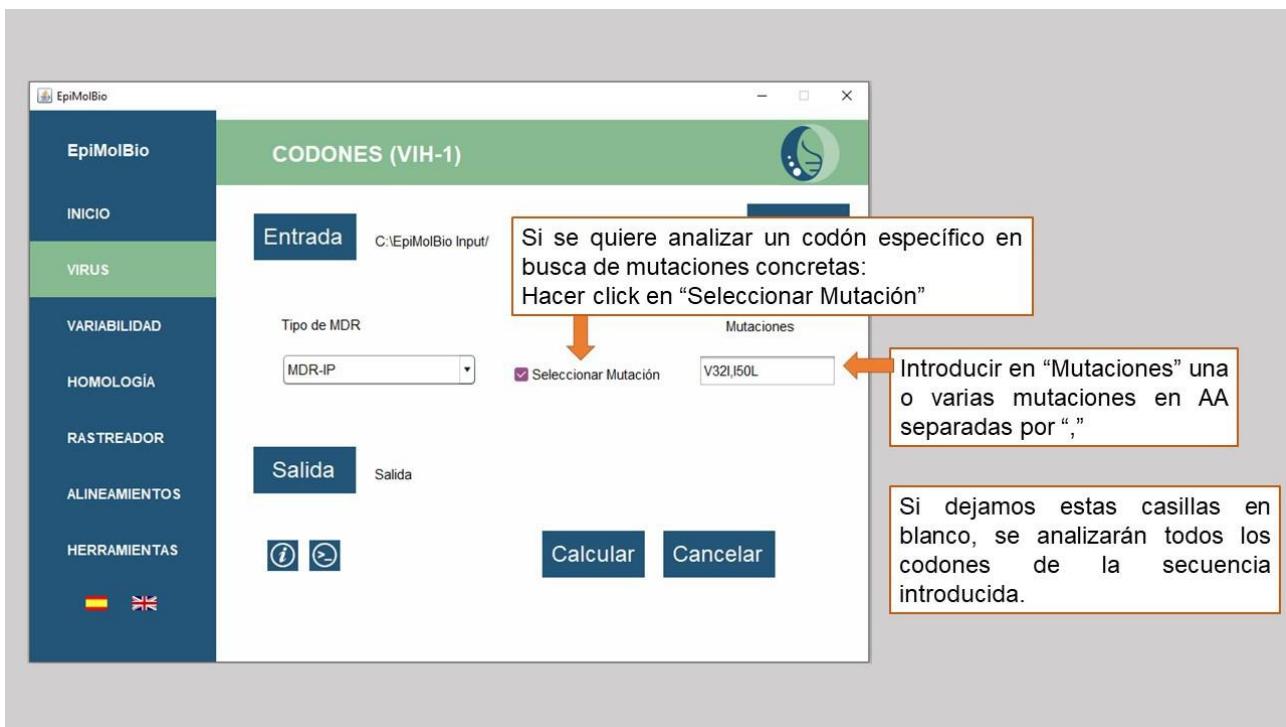
3)



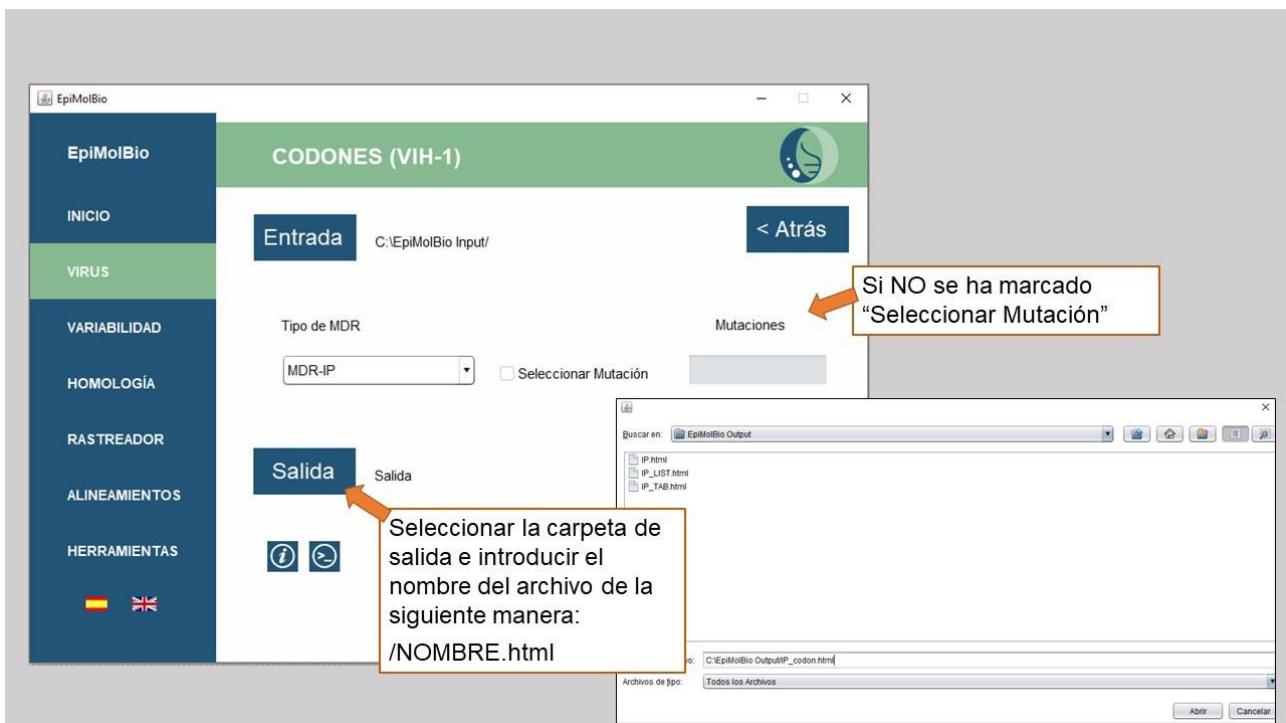
4)

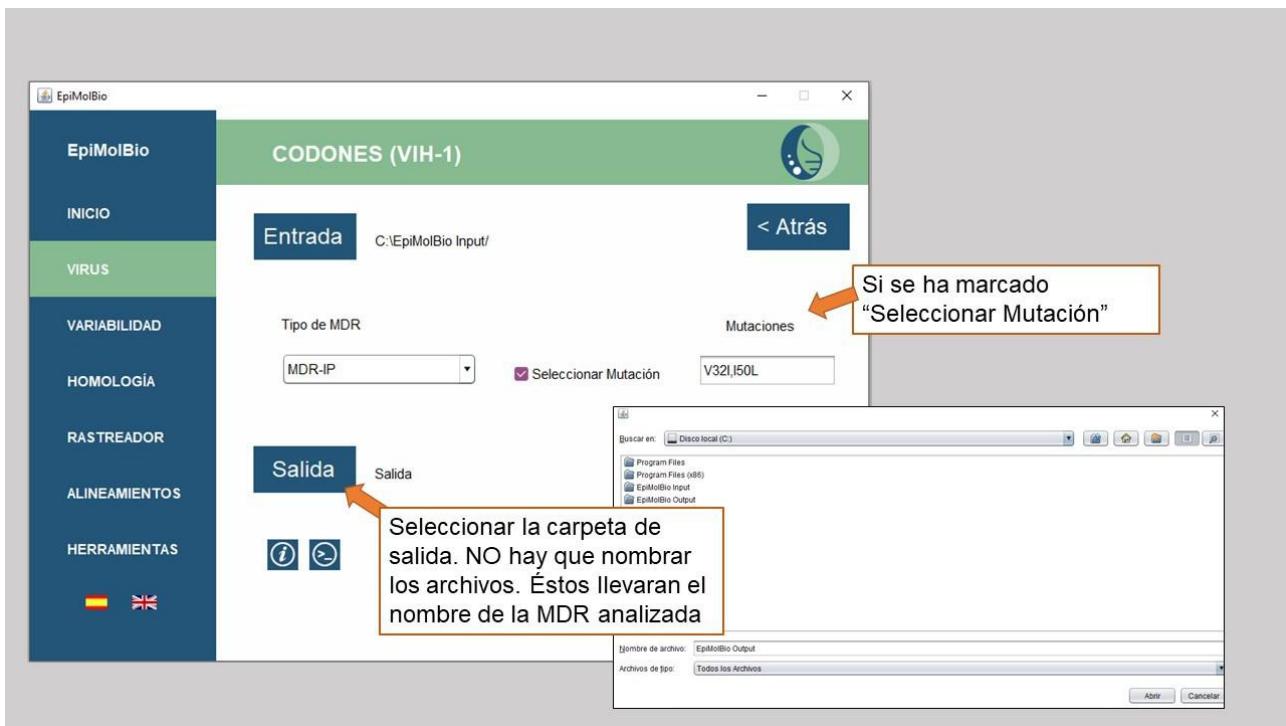


5)

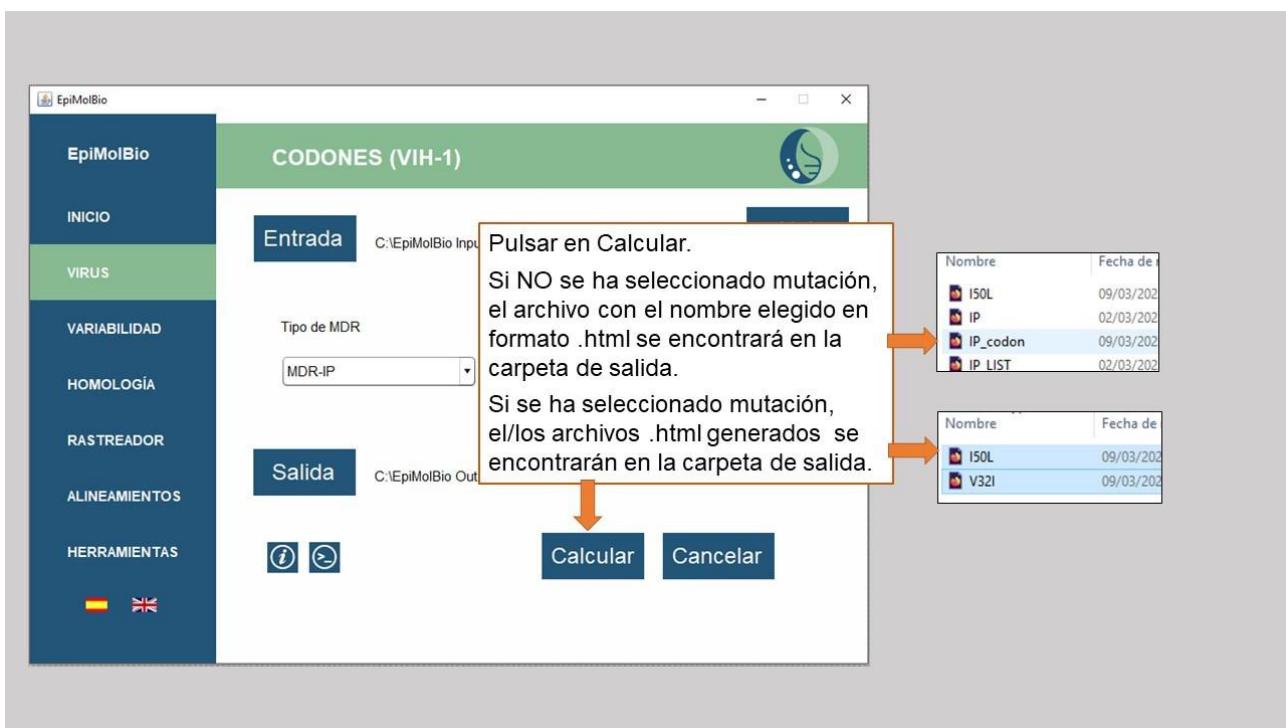


6)





7)



I.1.B) OTRAS MUTACIONES POL

Esta función permite detectar **cualquier mutación** (no sólo MDR) del VIH-1 o VIH-2 a partir de secuencias de las proteínas de Pol y obtener su porcentaje de aparición con respecto a la secuencia de referencia.

Se analizan las mutaciones o cambios de aminoácidos detectados en las proteínas Pol del VIH: Proteasa, Retrotranscriptasa o Integrasa. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis. EpiMolBio no detecta delecciones ni inserciones.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente las secuencias alineadas de la proteína de Pol que se quiera analizar (proteasa, retrotranscriptasa o integrasa) en aminoácidos y en formato .fasta. Esta carpeta puede contener un solo archivo o varios archivos .fasta si queremos analizar lo mismo en distintos grupos de secuencias (ej.: archivos divididos por variantes del VIH, por país de origen, año, etc.).

El archivo de **salida** será un archivo con extensión **.html**. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

En el campo “**Proteína**” habrá que seleccionar la proteína que se quiera analizar:

PR: Proteasa

RT: Retrotranscriptasa

IN: Integrasa

Luego, habrá que escoger el **tipo de VIH** que se va a analizar para establecer la secuencia de referencia: **VIH-1** o **VIH-2**. La secuencia de referencia para VIH-1 que emplea EpiMolBio es HXB2 (NCBI K03455.1) y para VIH-2 es ALI (NCBI AF082339).

En el campo “**Formato de Salida**” seleccionar el **formato de salida** deseado, pudiendo escoger entre tres tipos de formato de salida: **lista, tabla y tabla resumen**.

En el formato de salida “**Lista**” existen dos porcentajes de cribado distintos: 100% y >75%. Esto quiere decir que en el primero (100%) se muestran todas las mutaciones encontradas y en el segundo (>75%), las que tienen una frecuencia de aparición superior al 75%.

En los otros formatos de salida, “**Tabla**” y “**Tabla resumen**” el cribado es >75% por defecto.

En los todos los formatos se obtendrán las **mutaciones** detectadas en las secuencias introducidas en el archivo de entrada y su **porcentaje** con respecto al total de secuencias analizadas según el **código de colores** descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

1.- Lista:

En este formato de salida aparece, en la parte superior, el título del análisis y el cribado aplicado, seguido del nombre del archivo de entrada. En la columna “Posición” aparecen

todas las posiciones con su aminoácido de referencia. En la columna “Residuos” se muestran todos los residuos encontrados para esa posición según el cribado que se haya aplicado y su porcentaje de aparición coloreado según el código de colores. En la columna “Posiciones Totales” se describe el número total de secuencias válidas para esa posición que están presentes en el archivo analizado.

Ejemplo de formato de salida Lista para el análisis de Otras Mutaciones Pol con cribado al 100%:

| Lista Otras Mutaciones Pol Proteasa VIH-1 100% | | |
|--|--|--------------------|
| PR_procesado_traducido_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| P1 | P(99.896%) S(0.078%) A(0.004%) L(0.007%) T(0.007%) H(0.004%) V(0.004%) | 26838 |
| Q2 | Q(99.782%) E(0.071%) S(0.019%) H(0.056%) D(0.004%) K(0.023%) I(0.015%) P(0.008%) R(0.011%) T(0.004%) *(0.008%) | 26649 |
| V3 | I(99.858%) V(0.078%) N(0.015%) L(0.041%) T(0.007%) | 26831 |
| T4 | T(99.858%) M(0.004%) I(0.048%) N(0.019%) P(0.022%) S(0.034%) F(0.004%) A(0.007%) H(0.004%) | 26816 |
| L5 | L(99.888%) F(0.075%) V(0.015%) S(0.004%) R(0.007%) I(0.007%) T(0.004%) | 26780 |
| W6 | W(99.929%) G(0.030%) R(0.022%) *(0.007%) C(0.011%) | 26836 |

Ejemplo de formato de salida Lista para el análisis de Otras Mutaciones Pol con cribado >75%:

| Lista Otras Mutaciones Pol Proteasa VIH-1 > 75% | | |
|---|------------|--------------------|
| PR_procesado_traducido_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| V3 | I(99.858%) | 26831 |
| E35 | D(86.051%) | 26160 |
| M36 | I(99.177%) | 26743 |
| S37 | N(92.755%) | 26461 |
| R41 | K(97.554%) | 26572 |
| H69 | K(97.972%) | 26531 |
| L89 | M(96.625%) | 26547 |

| PR_procesado_traducido_02_AG.fasta | | |
|------------------------------------|------------|--------------------|
| Posición | Residuos | Posiciones Totales |
| V3 | I(99.529%) | 9557 |
| I13 | V(91.362%) | 9308 |
| K20 | I(94.888%) | 9448 |
| M36 | I(98.475%) | 9511 |
| R41 | K(92.213%) | 9374 |
| H69 | K(96.945%) | 9394 |
| L89 | M(92.325%) | 9407 |

2.- Tabla:

En el formato de salida Tabla se aplica por defecto el cribado >75%, por lo que las mutaciones con una frecuencia ≤ 75% no aparecerán en este formato.

En este formato de salida aparece, en la parte superior, el título del análisis. En la primera columna, “Archivo”, se muestran los nombres de los archivos de entrada empleados para generar la tabla. En las siguientes columnas, se muestran cada una de las posiciones con su aminoácido de referencia y el residuo mutado con la celda coloreada según el código de colores, indicando el porcentaje de aparición para esa posición.

Ejemplo de formato de salida Tabla para el análisis de Otras Mutaciones Pol:

| Archivo | | | | | | | | | | | | | | | | | | | | | |
|-------------------------------------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| | P1 | Q2 | V3 | T4 | L5 | W6 | Q7 | R8 | P9 | L10 | V11 | T12 | I13 | K14 | I15 | G16 | G17 | Q18 | L19 | K20 | E21 |
| PR_procesado_traducido_01_AE.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_02_AG.fasta | | | I | | | | | | | | | | V | | | | | | | I | |
| PR_procesado_traducido_03_A6B.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_04_cpx.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_05_DF.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_06_cpx.fasta | | | I | | | | | | | | | | V | | | | | | | I | |
| PR_procesado_traducido_07_BC.fasta | | | I | | | | | | | | | | | | | | | | | | |
| PR_procesado_traducido_08_BC.fasta | | | I | | | | | | | | | S | | | V | | | | I | | |
| PR_procesado_traducido_09_cpx.fasta | | | I | | | | | | | | | | V | | | | | | | | |

3.- Tabla Resumen:

En el formato de salida Tabla Resumen se aplica por defecto el cribado >75% por lo que las mutaciones con una frecuencia ≤ 75% no aparecerán en este formato.

En este formato de salida aparece, en la parte superior, el título del análisis. En la primera columna, “Archivo”, se muestran los nombres de los archivos de entrada empleados para generar la tabla. En la siguiente columna, “Residuos”, aparecen cada una de las posiciones con su aminoácido de referencia y el residuo mutado coloreado según el código de colores que indica el porcentaje de aparición para esa posición. En la columna “Secuencias Totales” aparece el número de secuencias totales de cada archivo de entrada.

Ejemplo de formato de salida Tabla Resumen para el análisis de Otras Mutaciones Pol:

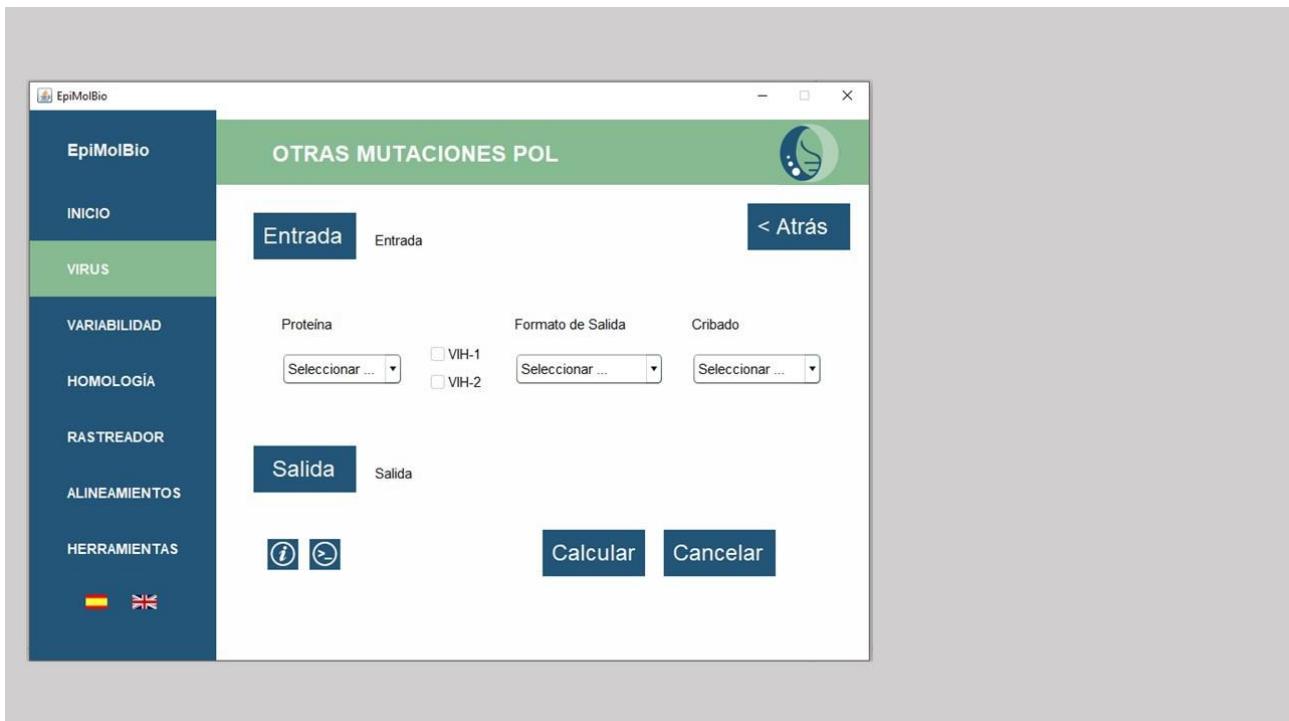
| Tabla Resumen Otras Mutaciones Pol Proteasa VIH-1 > 75% | | |
|---|--|--------------------|
| Archivo | Residuos | Secuencias Totales |
| PR_procesado_traducido_01_AE.fasta | V3I, E35D, M36I, S37N, R41K, H69K, L89M | 26849 |
| PR_procesado_traducido_02_AG.fasta | V3I, I13V, K20I, M36I, R41K, H69K, L89M | 9577 |
| PR_procesado_traducido_03_A6B.fasta | V3I, E35D, M36I, S37N, R41K, H69K, L89M | 310 |
| PR_procesado_traducido_04_cpx.fasta | V3I, M36I, R41K, H69K | 15 |
| PR_procesado_traducido_05_DF.fasta | V3I, S37N, R41K | 24 |

Paso a paso:

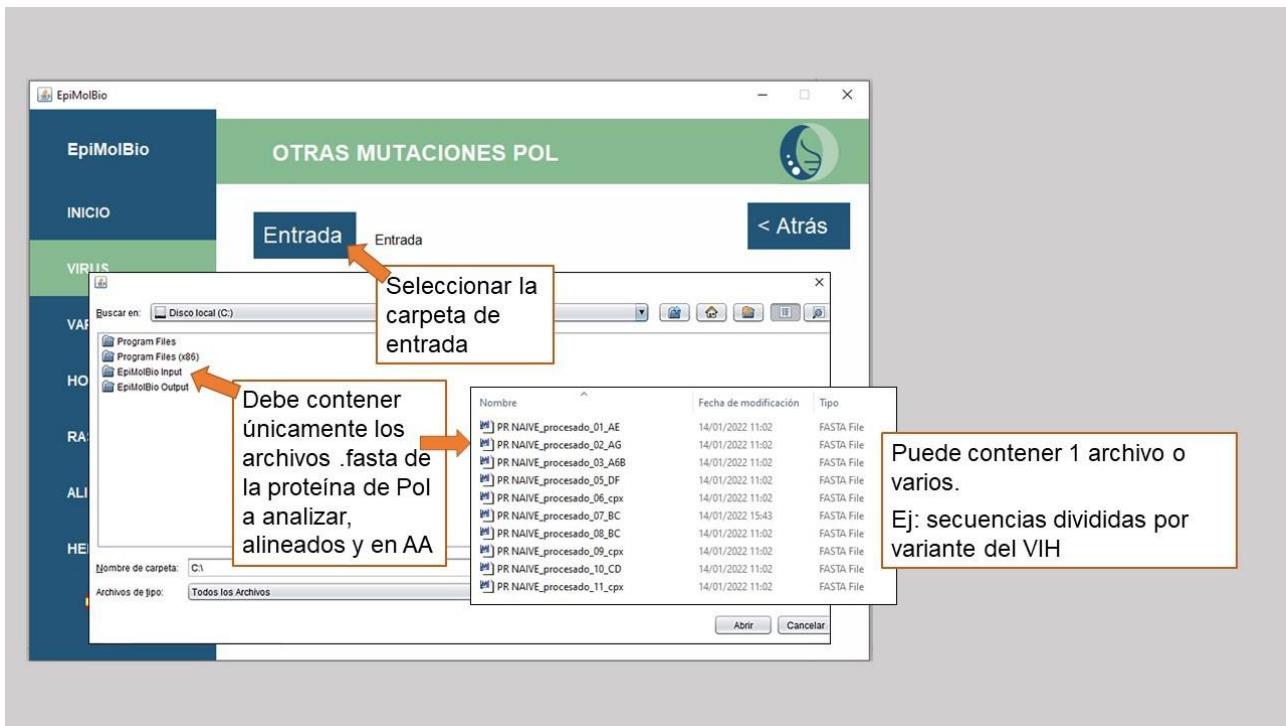
1)



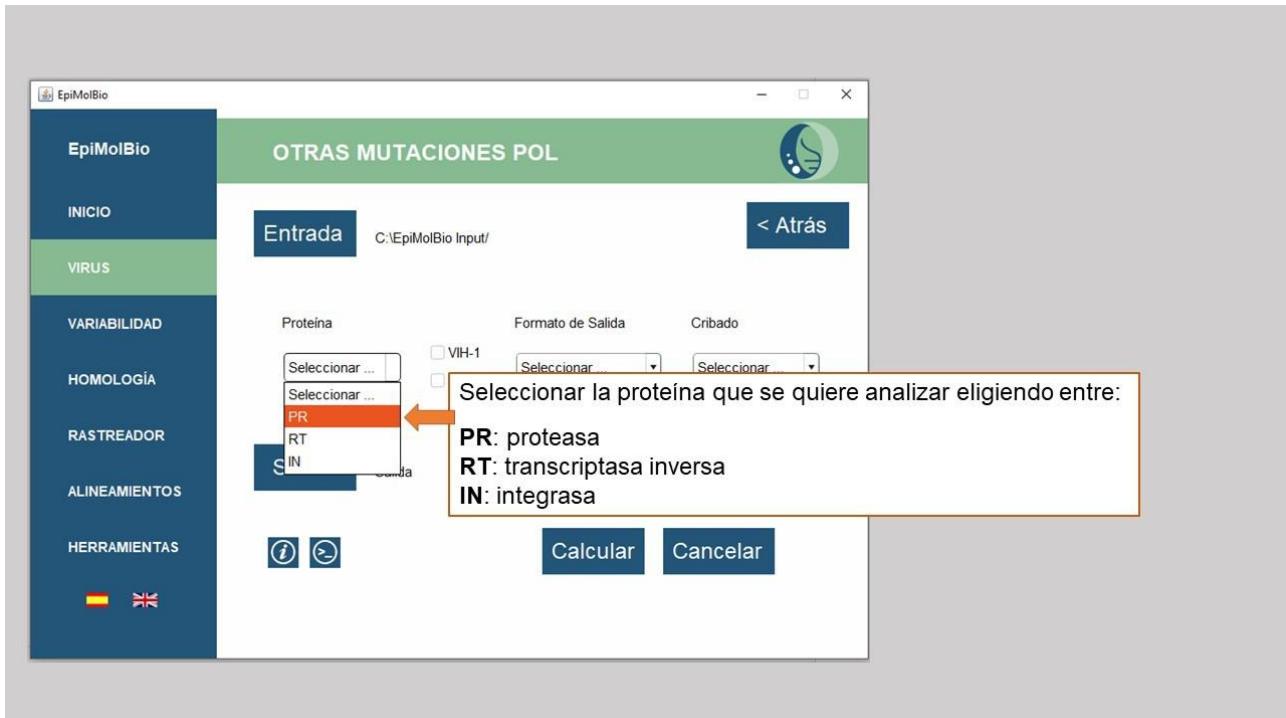
2)



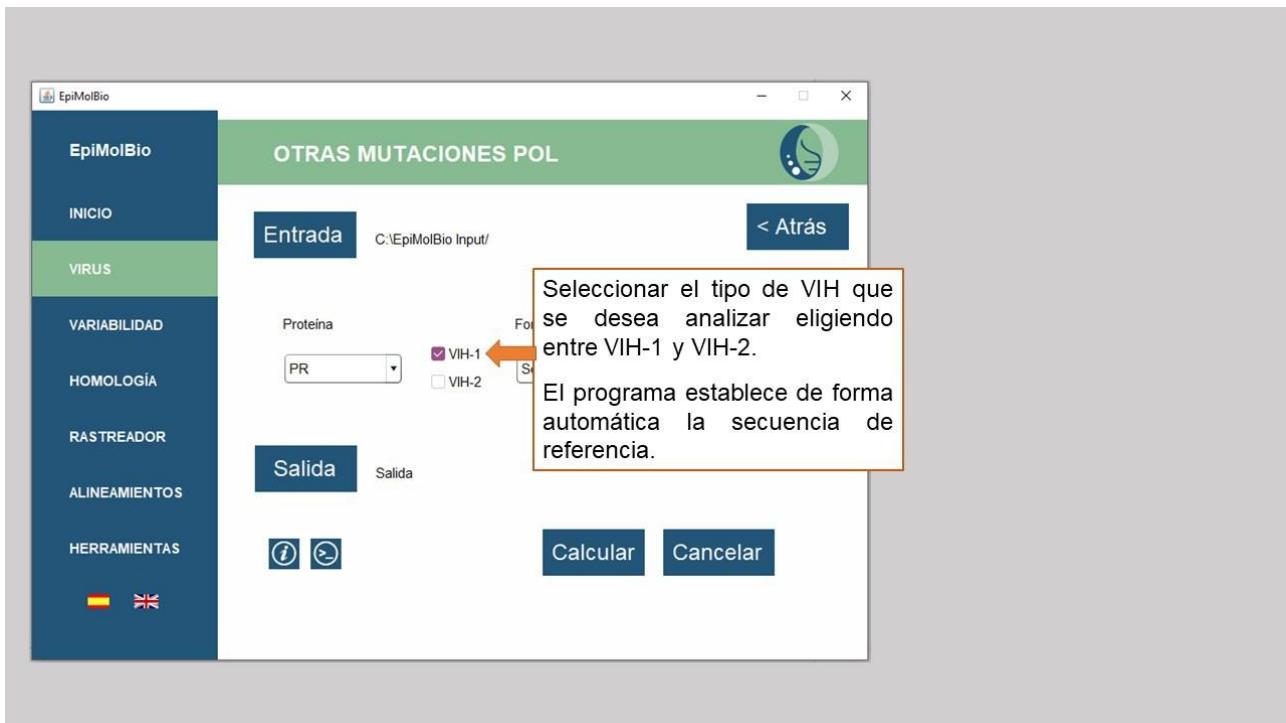
3)



4)



5)



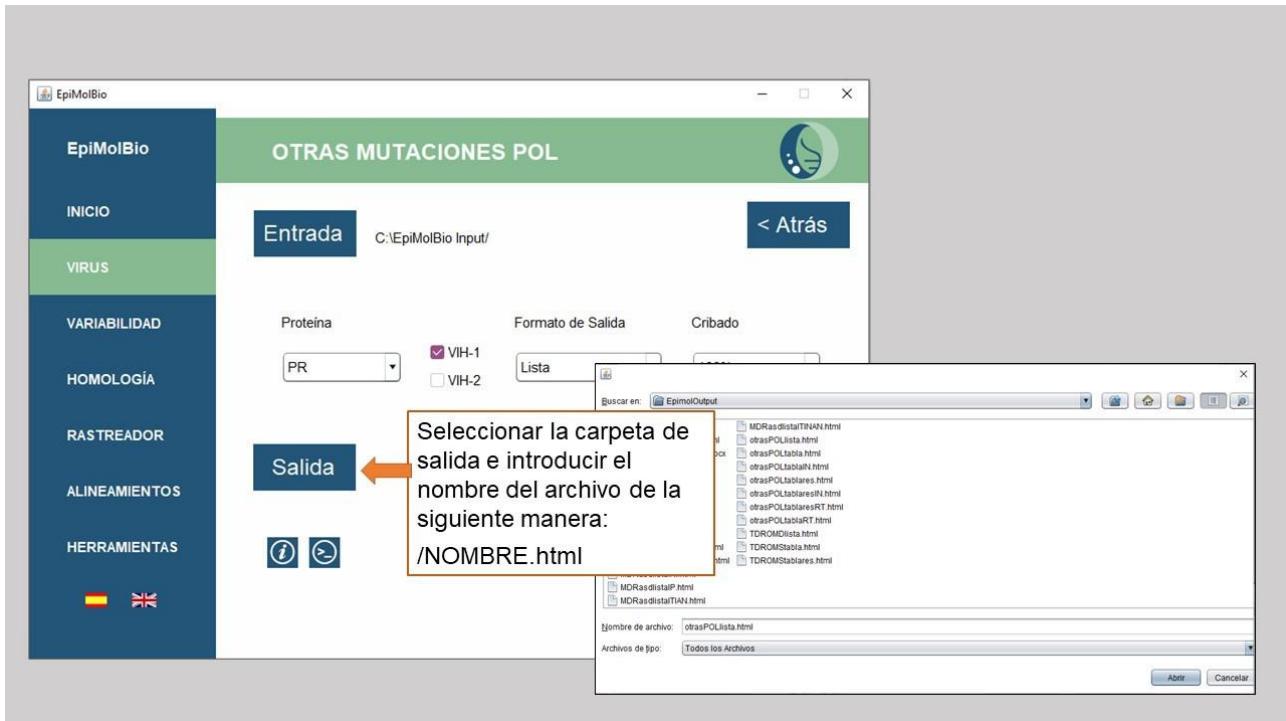
6)



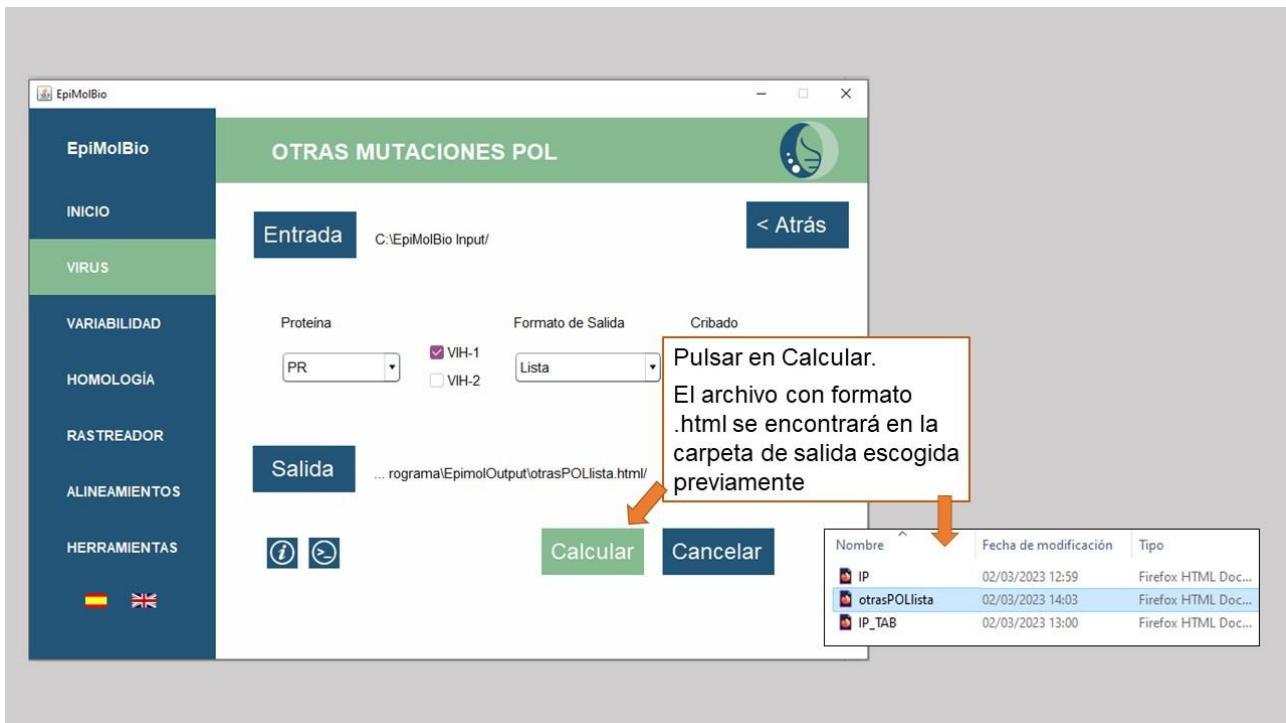
7)



8)



9)



I.1.C) CONSERVACIÓN POL

Esta función genera una tabla .html donde se muestra el consenso de las secuencias de entrada con el aminoácido más prevalente en cada posición de la secuencia de aminoácidos de la proteína Pol seleccionada, permitiendo conocer el residuo más conservado de la proteína para cada posición. También muestra una tabla con los residuos que presentan una conservación superior al 75% para cada posición. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis. Esta función sirve para las proteínas del gen *pol* del VIH-1 y del VIH-2.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente las secuencias alineadas de la proteína de Pol que se quiera analizar (proteasa, retrotranscriptasa o integrasa) en aminoácidos y en formato .fasta. Esta carpeta puede contener un solo archivo o varios archivos .fasta si queremos analizar lo mismo en distintos grupos de secuencias (ej.: archivos divididos por variantes del VIH).

En el campo “**Proteína**” habrá que seleccionar la proteína que se quiera analizar:

PR: Proteasa

RT: Retrotranscriptasa o Transcriptasa Inversa

IN: Integrasa

Seleccionar el **tipo de VIH** que se va a analizar para establecer la longitud de cada proteína según la secuencia de referencia: **VIH-1** (HXB2, NCBI K03455.1) o **VIH-2** (ALI, NCBI AF082339).

El archivo de **salida** será un archivo con extensión **.html**. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

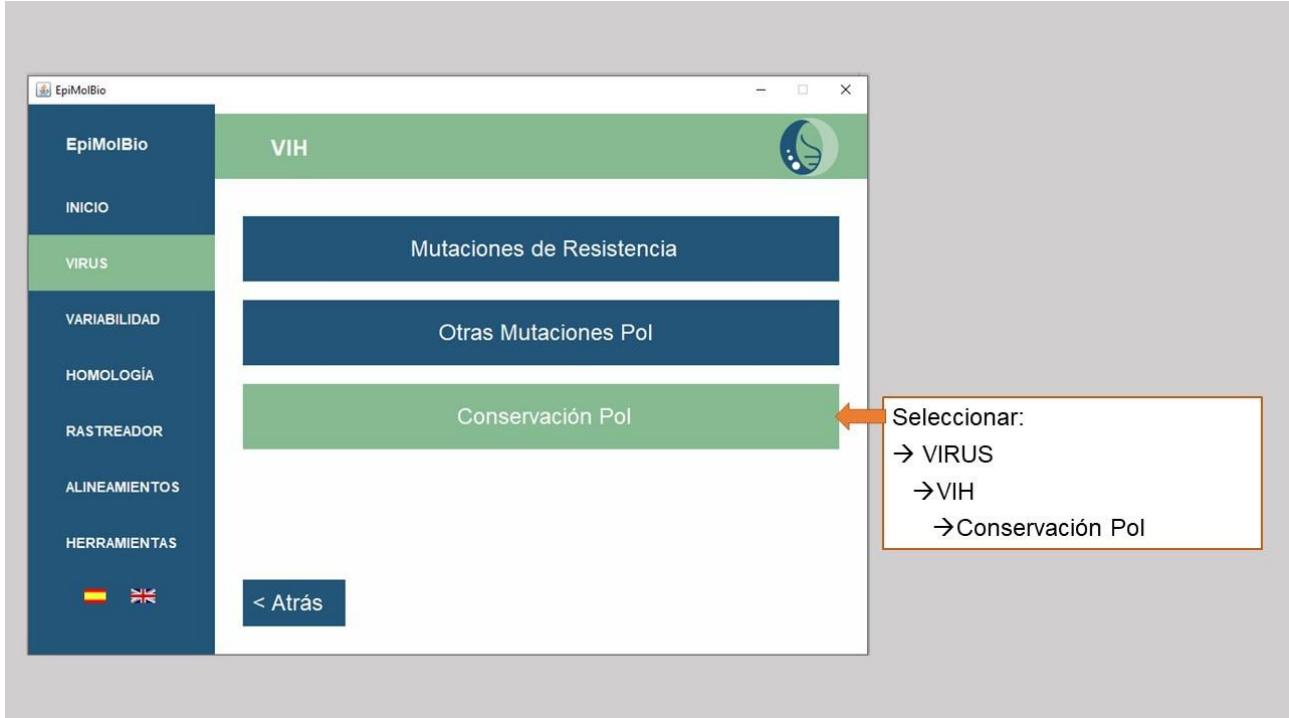
En el archivo de salida aparece, en la parte superior, el título del análisis seguido del nombre del archivo de entrada. Debajo aparece la secuencia consenso para cada archivo con los residuos coloreados según su porcentaje siguiendo el código de colores descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul. Cuando se detecte más de un residuo con el mismo porcentaje en la misma posición, éstos aparecerán entre paréntesis. Debajo, en la columna “Posición” aparece cada una de las posiciones con su aminoácido de referencia. En la columna “Residuos” aparece el aminoácido más frecuente para cada posición, seguido de su porcentaje coloreado según el código de colores siempre que esté presente en una frecuencia superior al 75%. En la columna “Posiciones Totales” aparece el número total de residuos válidos para esa posición.

Ejemplo de formato de salida para el análisis de Conservación Pol:

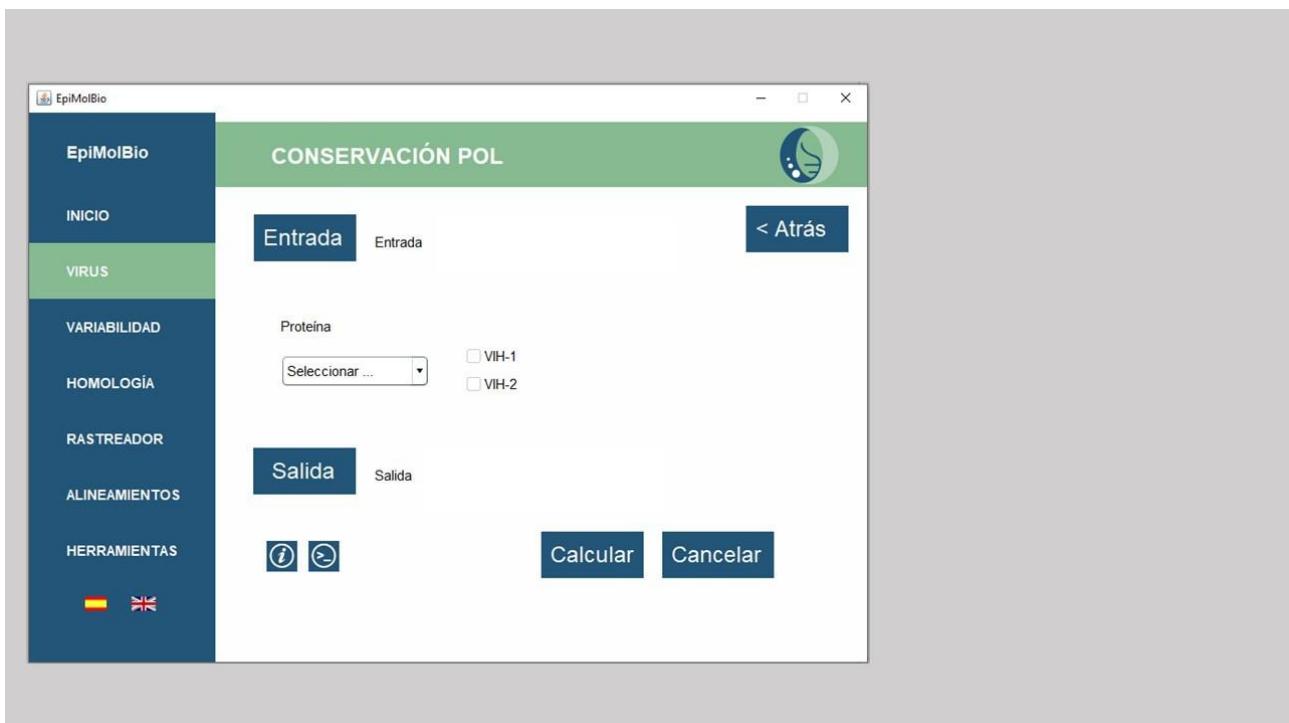
| Conservación Pol Proteasa VIH-1 > 75% | | |
|---------------------------------------|--|--------------------|
| PR_procesado_traducido_01_AE.fasta | | |
| CONSENSO | PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIVRQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQ/GCTLNF | |
| Posición | Residuos | Posiciones Totales |
| P1 | P(99.896%) | 26838 |
| Q2 | Q(99.782%) | 26649 |
| V3 | I(99.858%) | 26831 |
| T4 | T(99.858%) | 26816 |
| L5 | L(99.888%) | 26780 |
| W6 | W(99.929%) | 26836 |
| Q7 | Q(99.751%) | 26536 |
| R8 | R(99.922%) | 26792 |
| P9 | P(99.955%) | 26613 |

Paso a paso:

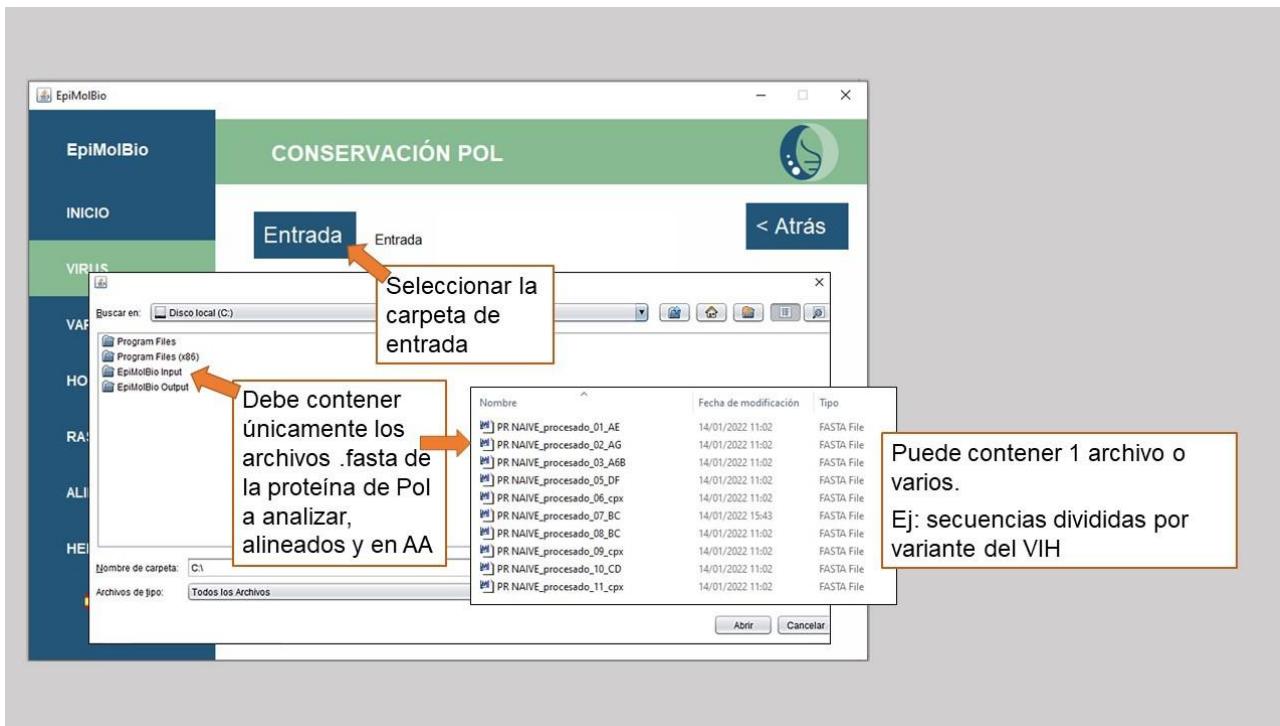
1)



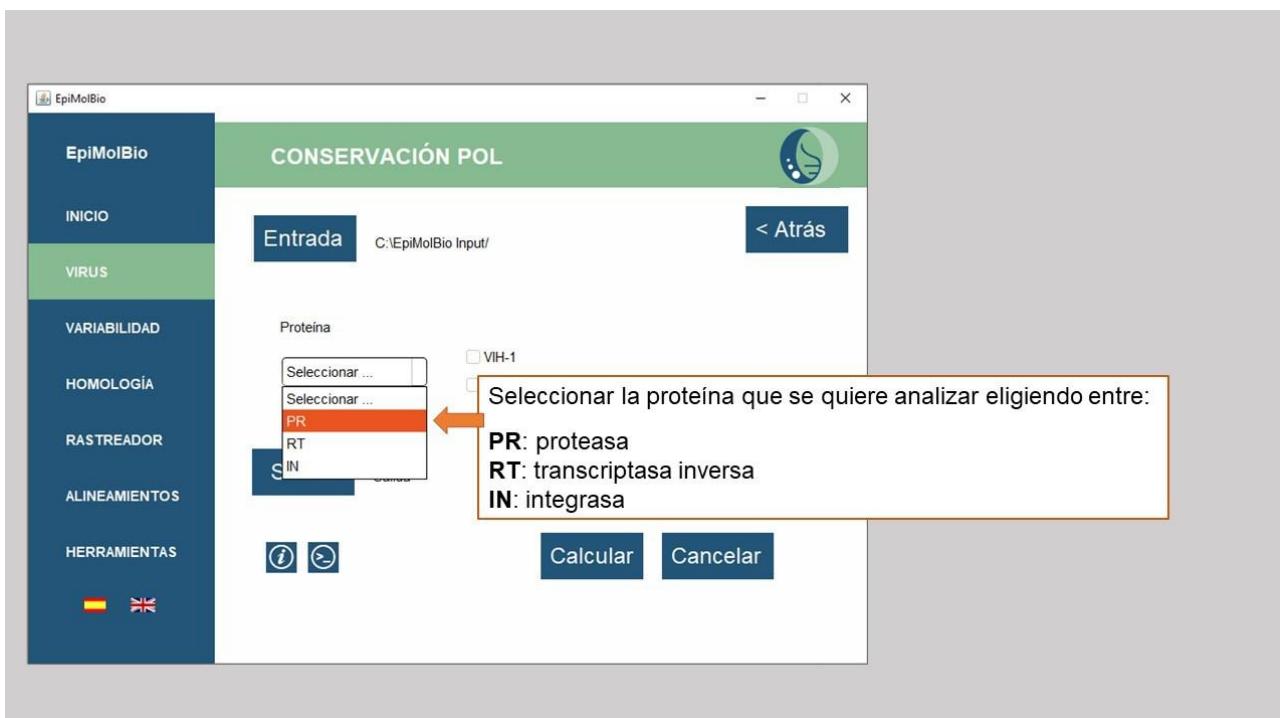
2)



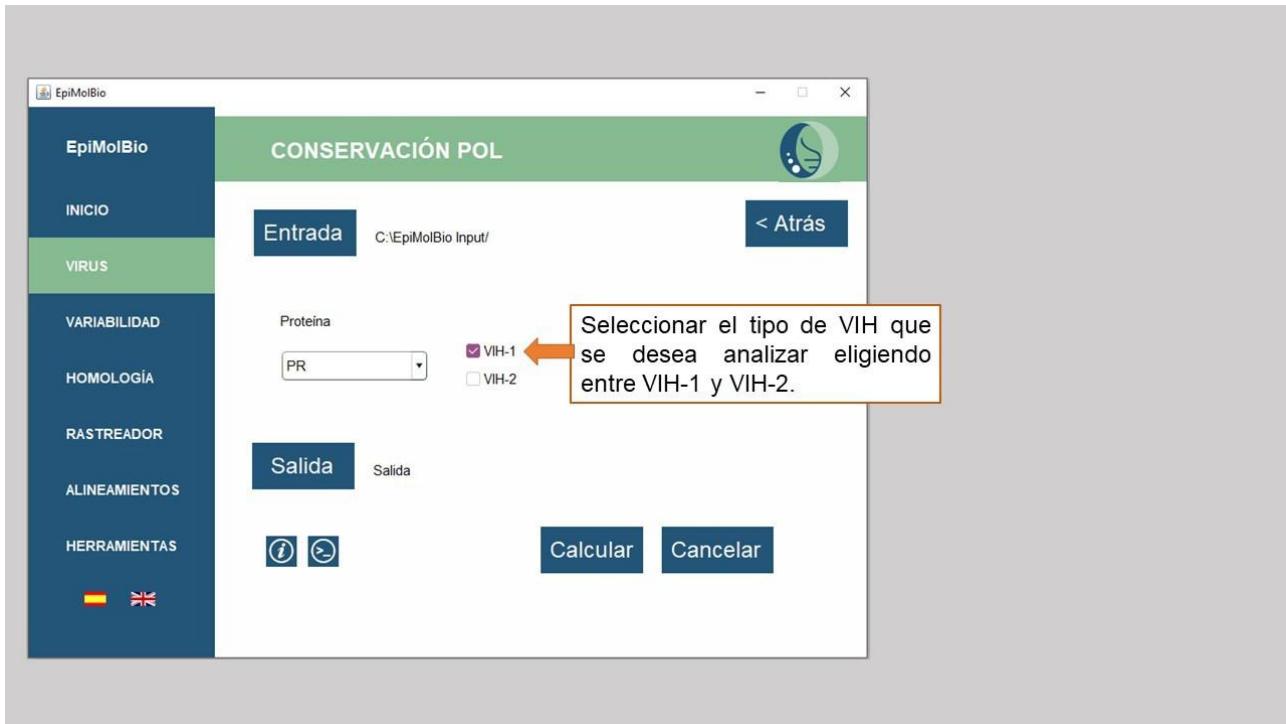
3)



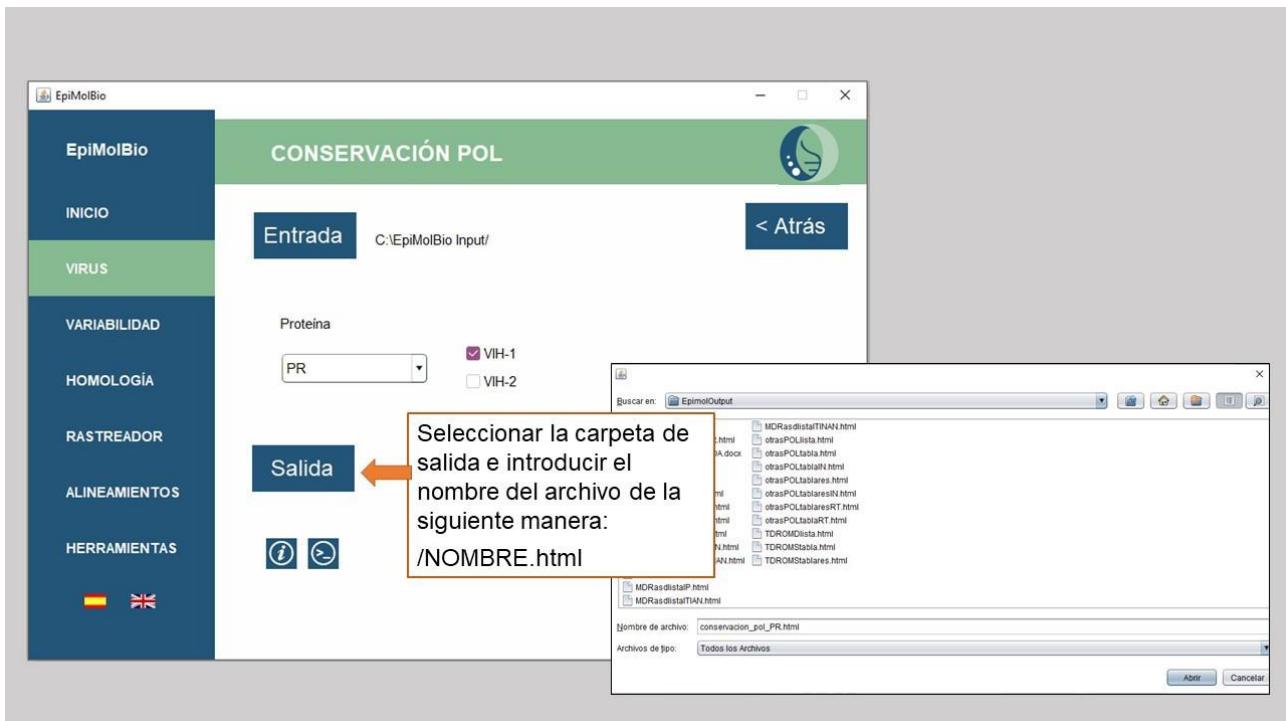
4)



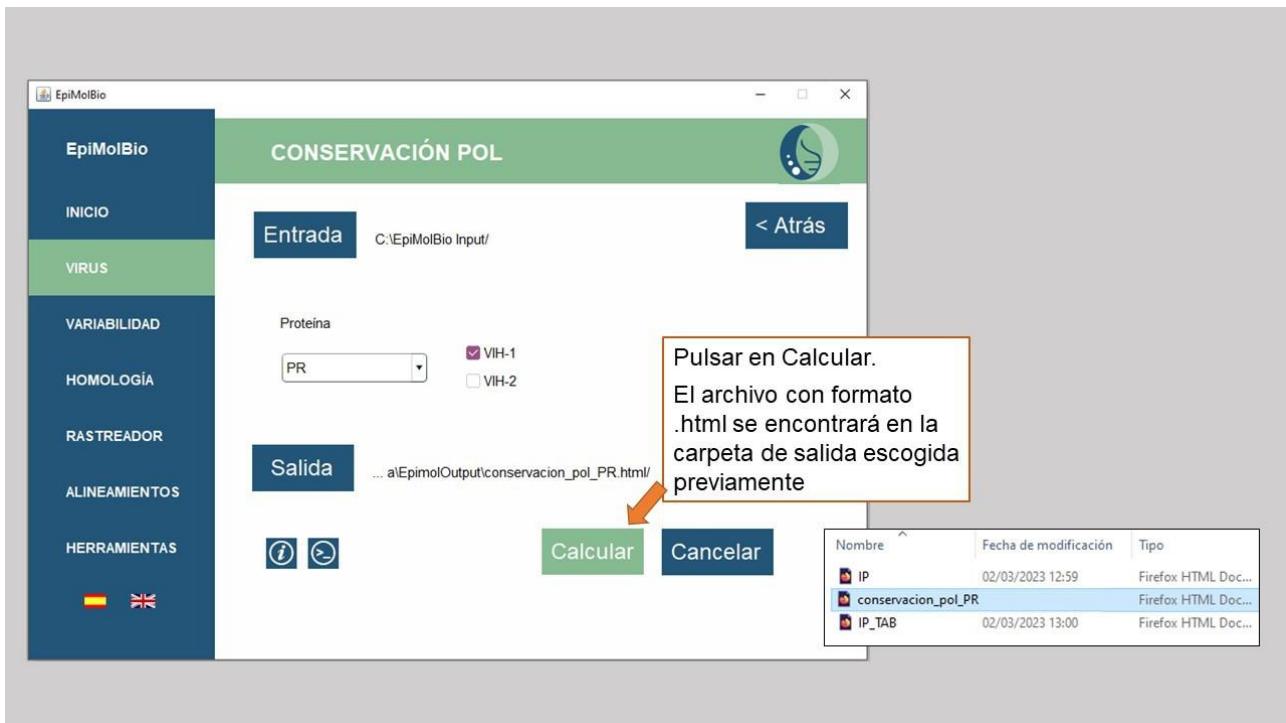
5)



6)



7)



I.2. SARS-CoV-2

RASTREADOR DE PROTEÍNAS

Esta función genera archivos .fasta con las secuencias de nucleótidos o de aminoácidos de las proteínas que seleccionemos del SARS-CoV-2 a partir de genomas completos exclusivamente.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con los genomas completos del SARS-CoV-2 en nucleótidos. EpiMoBio rastrea las proteínas dentro del rango de posiciones en el que se encuentran según la secuencia de referencia Wuhan (NC 045512.2).

En el campo “**Seleccionar Proteína**” habrá que seleccionar la proteína del SARS-CoV-2 que se quiera rastrear, o bien seleccionar “Todas” para que se rastreen todas las proteínas.

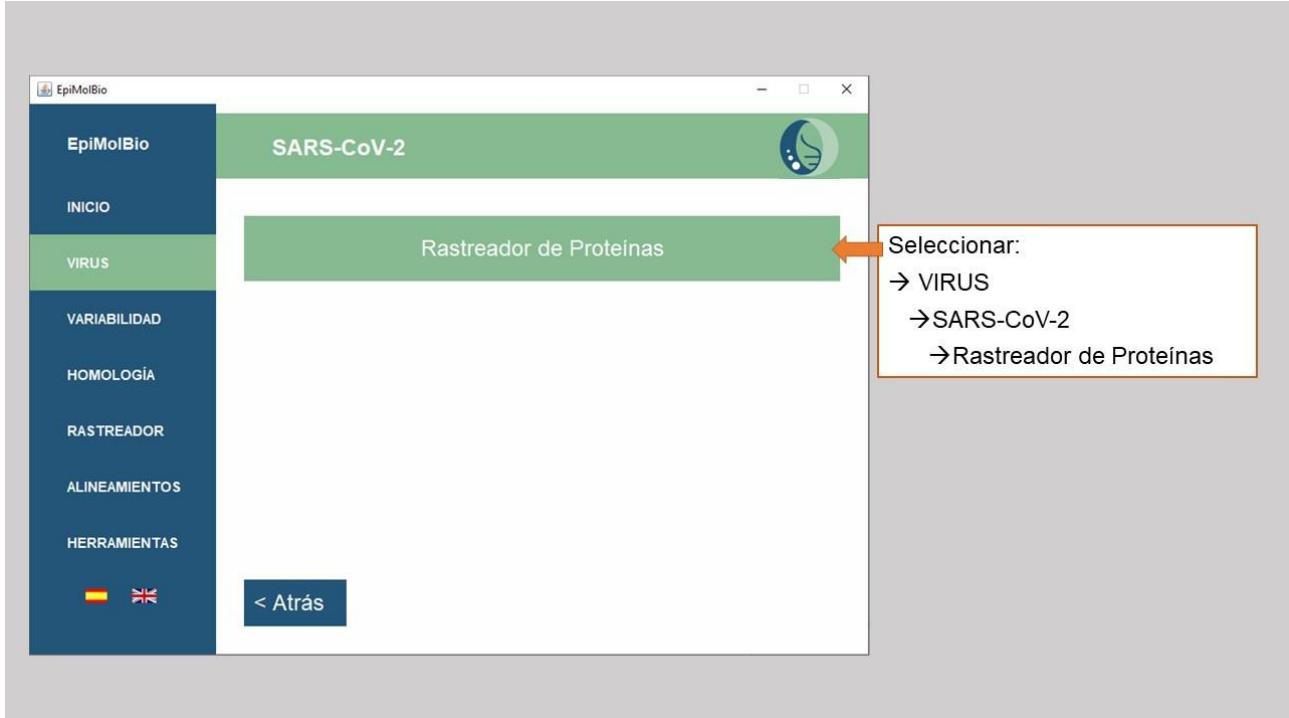
Escoger si se quiere el archivo .fasta de salida en nucleótidos (escoger **No Traducir**) o traducido en aminoácidos (escoger **Traducir**).

El archivo de **salida** será un archivo tipo .fasta. Por cada archivo de la carpeta de entrada, se obtiene uno de salida con las secuencias encontradas en nucleótidos o traducidas a aminoácidos. Puede que el número de secuencias de entrada no coincida con el número de proteínas rastreadas. Cuando una secuencia de entrada contiene muchas mutaciones o está incompleta en la región que corresponde a la proteína buscada, el Rastreador de Proteínas no puede reconocerla y, por lo tanto, el programa no la recuperará. La eficacia del Rastreador de Proteínas se resume en el **Anexo II**.

Seleccionar la **carpeta de salida** donde queremos que aparezcan los archivos .fasta. El archivo se nombrará automáticamente de la siguiente manera: Proteína seleccionada_Rastreado_Nombre del archivo de entrada.fasta (ej.: S (Spike)_Rastreado_sequences.fasta). Si se selecciona “Todas” en “Seleccionar Proteína”, cada proteína estará separada en un archivo .fasta con el nombre correspondiente.

Paso a paso

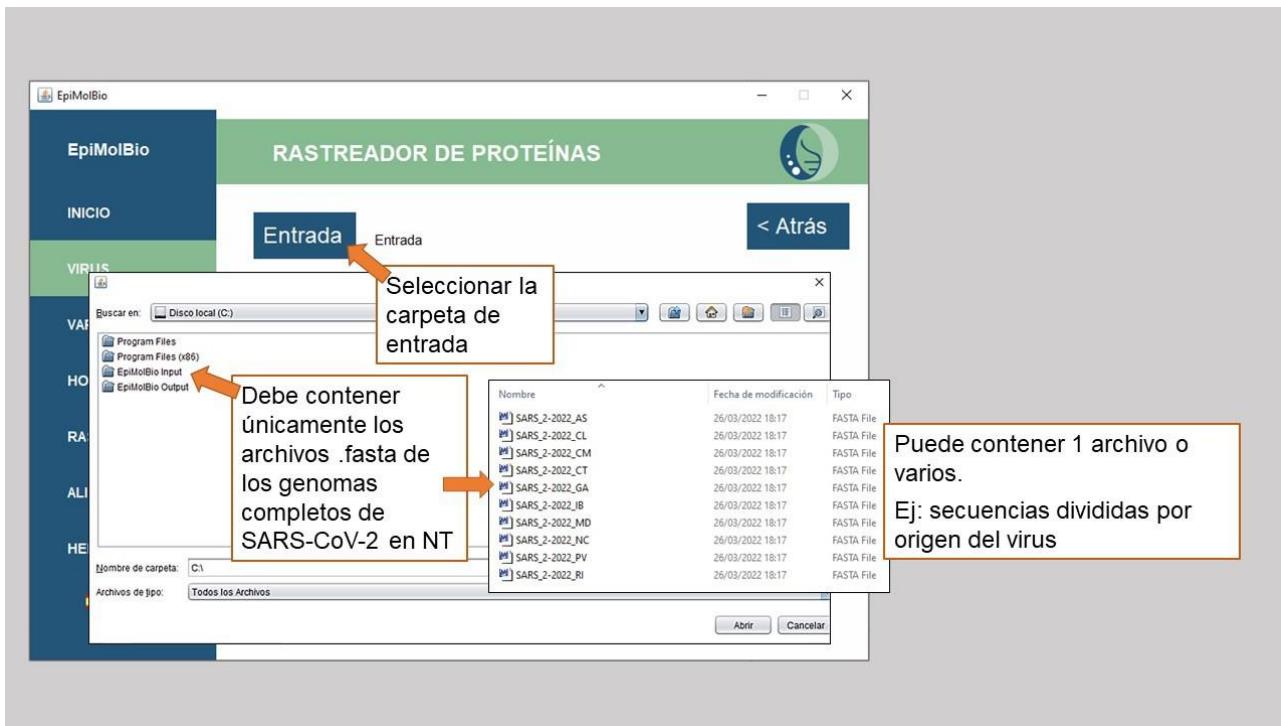
1)



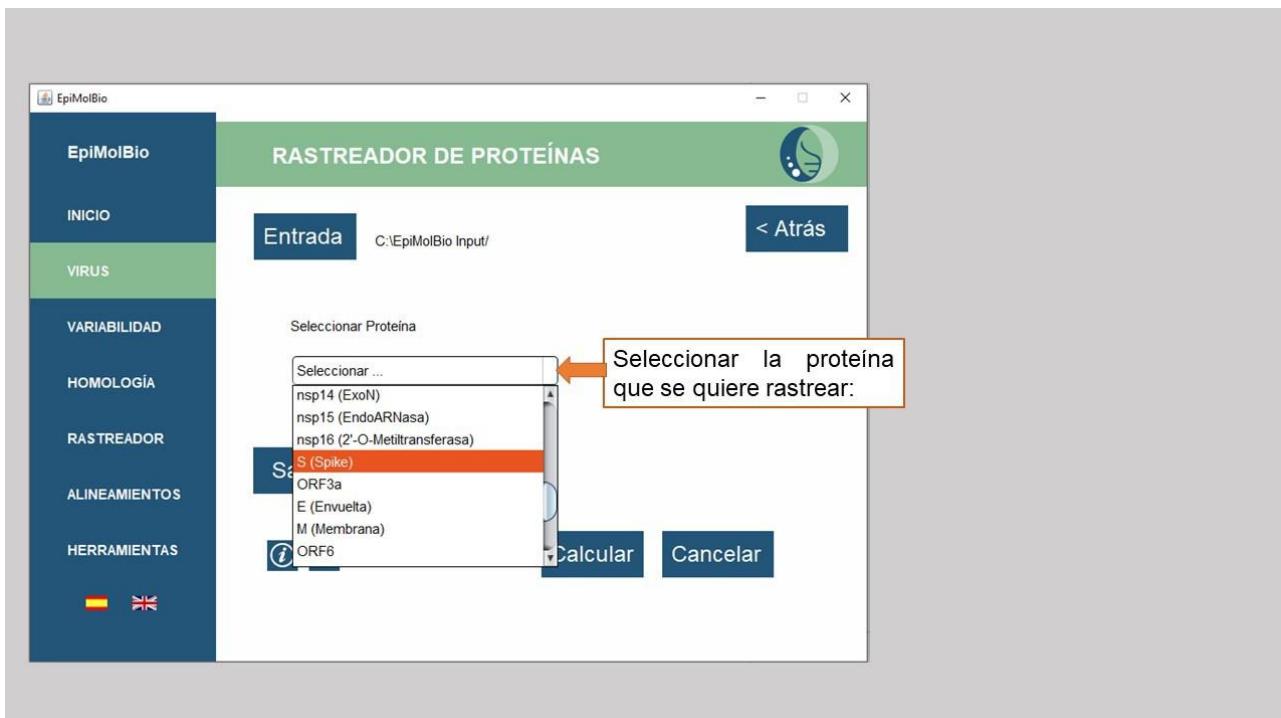
2)



3)



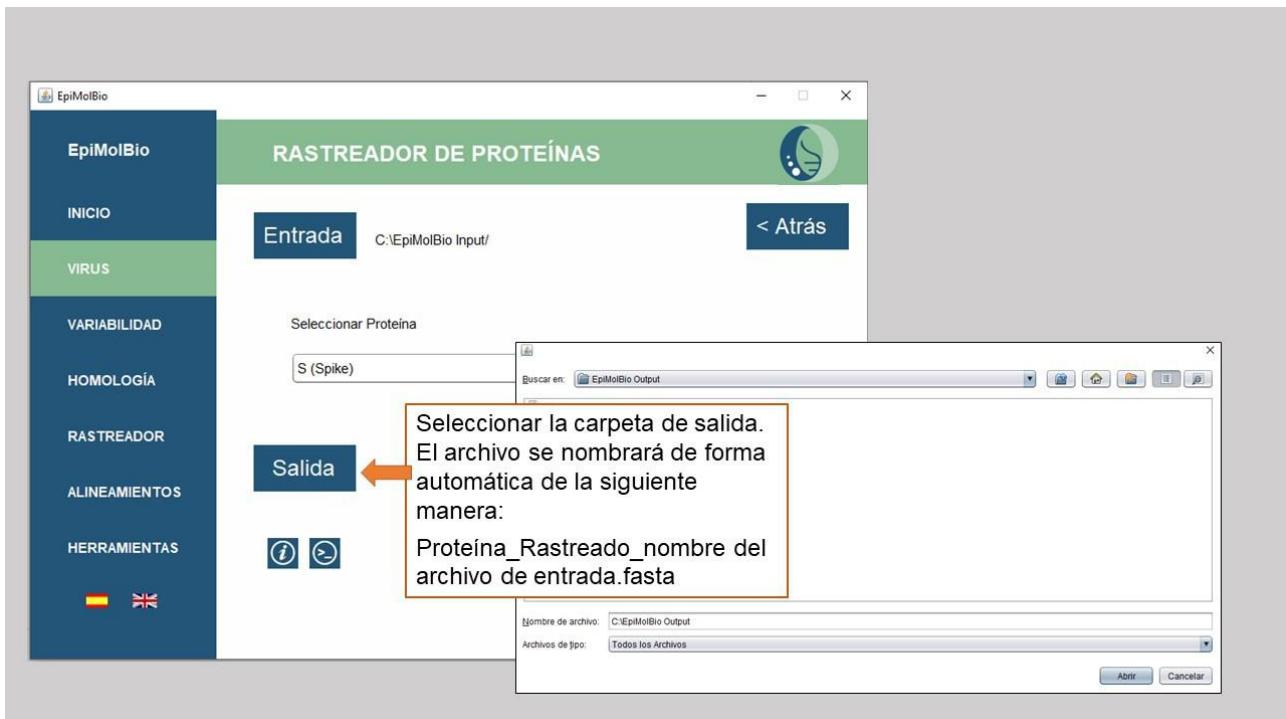
4)



5)



6)



7)



II. VARIABILIDAD

En esta sección se encuentran varias herramientas específicas para el análisis de variabilidad de cualquier secuencia genética.

II.1. POLIMORFISMOS

Permite la detección de polimorfismos informando de su localización y frecuencia de aparición. La búsqueda puede realizarse analizando cada residuo de forma individual (II.1.A) o analizando codones (II.1.B).

II.1.A) INDIVIDUAL

Permite obtener las frecuencias de aparición de mutaciones utilizando como referencia cualquier secuencia introducida por el usuario. Estos análisis se pueden realizar empleando secuencias de aminoácidos o de nucleótidos. Para realizar el análisis en secuencias de nucleótidos será necesario emplear la herramienta **Buscar y Reemplazar** de **Edición de Archivos** y sustituir las “N” por “?” para que estas se excluyan del análisis, tal y como se explica en el apartado de “**Herramientas**”.

1.-Posiciones Mutadas:

Permite **detectar, localizar y conocer la frecuencia de aparición de mutaciones en una o varias posiciones** con respecto a una secuencia de referencia introducida por el usuario. Con esta herramienta se puede escoger buscar sólo una o varias posiciones concretas, así como seleccionar la frecuencia mínima de aparición de las mutaciones que queremos que se muestren. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis.

En el archivo de salida aparece, en la parte superior, el título del análisis seguido del nombre del archivo de entrada. Debajo, en la columna “Posición” aparece cada una de las posiciones con su aminoácido o nucleótido de referencia según la secuencia de referencia introducida por el usuario. En la columna “Residuos” aparece el residuo encontrado junto a su porcentaje de aparición coloreado según el código de colores descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul. Si se escoge el campo “Todas las Posiciones”, también aparecerá el porcentaje del residuo de referencia. En la columna “Posiciones Totales” aparece el número total de secuencias válidas para esa posición.

Ejemplo de formato de salida Posiciones Mutadas con Valor Mínimo 0.0 seleccionando Todas las Posiciones:

| Variabilidad Polimorfismos Individual Todas las Posiciones | | |
|--|--|--------------------|
| PR_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| P1 | P(99.896%) S(0.078%) A(0.004%) L(0.007%) T(0.007%) H(0.004%) V(0.004%) | 26838 |
| Q2 | Q(99.782%) E(0.071%) S(0.019%) H(0.056%) D(0.004%) K(0.023%) L(0.015%) P(0.008%) R(0.011%) T(0.004%) *(0.008%) | 26649 |
| V3 | I(99.858%) V(0.078%) N(0.015%) L(0.041%) T(0.007%) | 26831 |
| T4 | T(99.858%) M(0.004%) I(0.048%) N(0.019%) P(0.022%) S(0.034%) F(0.004%) A(0.007%) H(0.004%) | 26816 |
| L5 | L(99.888%) F(0.075%) V(0.015%) S(0.004%) R(0.007%) I(0.007%) T(0.004%) | 26780 |
| W6 | W(99.929%) G(0.030%) R(0.022%) *(0.007%) C(0.011%) | 26836 |

Para realizar este análisis, en **entrada** se debe seleccionar una **carpeta** donde se tengan exclusivamente los archivos en formato .fasta de las secuencias a analizar en nucleótidos o aminoácidos.

Habrá que seleccionar “**Posiciones Mutadas**” en el desplegable “**Formato de salida**”.

En “**Referencia**” introducir la secuencia de referencia en letras sin saltos de línea, teniendo en cuenta que la secuencia debe introducirse en nucleótidos o aminoácidos, según si los archivos de entrada están sin traducir o traducidos.

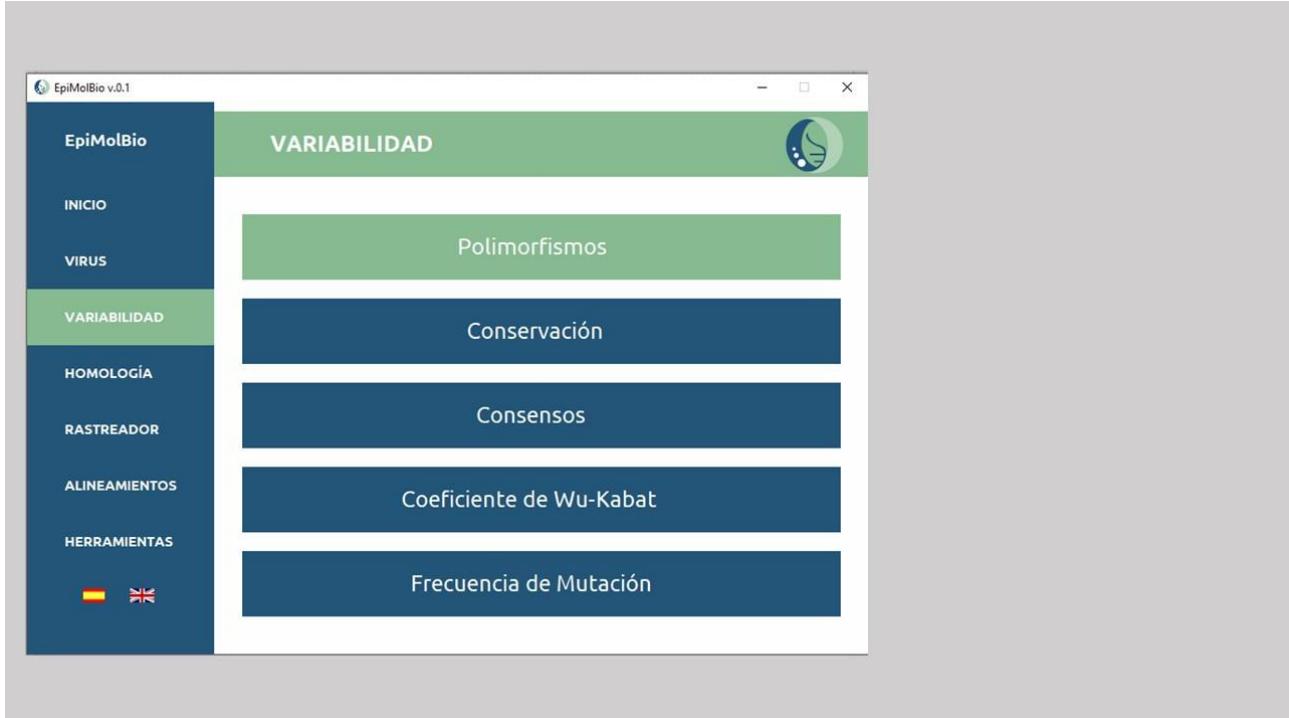
El campo “**Valor Mínimo**” puede dejarse vacío si se desea que el resultado muestre todas las mutaciones detectadas, independientemente de su frecuencia de aparición. Si se quiere filtrar sólo mutaciones que aparezcan en cierta frecuencia, habrá que introducir el valor mínimo en formato numérico decimal (ej.: 90.0 para que aparezcan sólo las mutaciones que se encuentran en un porcentaje mayor al 90%).

El campo “**Mutaciones**” puede dejarse vacío si se desea que el resultado muestre todas las mutaciones detectadas. Si se quiere buscar una o varias mutaciones concretas, habrá que introducir la posición del residuo donde se busca la mutación (ej.: 2). Cuando los archivos de entrada y la referencia estén en nucleótidos, introducir la posición que corresponde a la mutación buscada en nucleótidos (ej.: 6). Si se quieren buscar varias mutaciones, separarlas por una coma “,” sin espacios (ej.: 6,8,11).

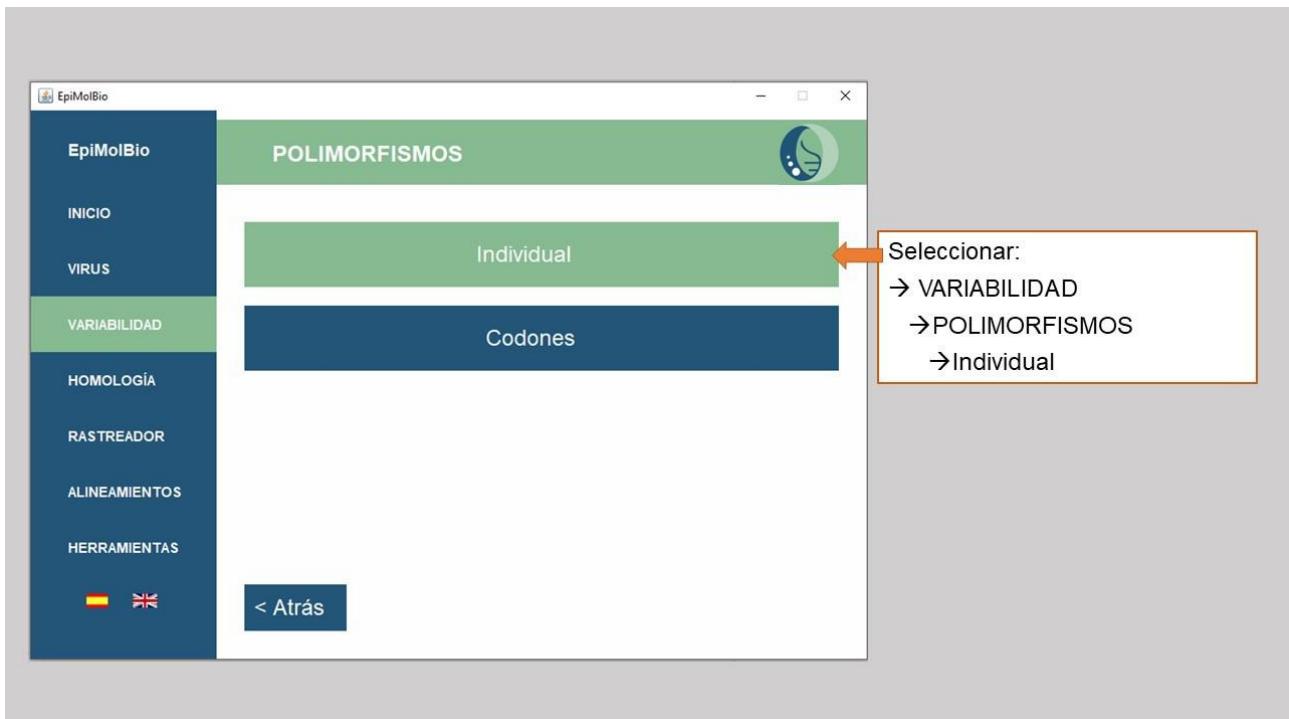
El resultado se muestra en un archivo .html. En **salida** se debe seleccionar la carpeta donde se quiera guardar el resultado y nombrar el archivo con la extensión .html.

Paso a paso:

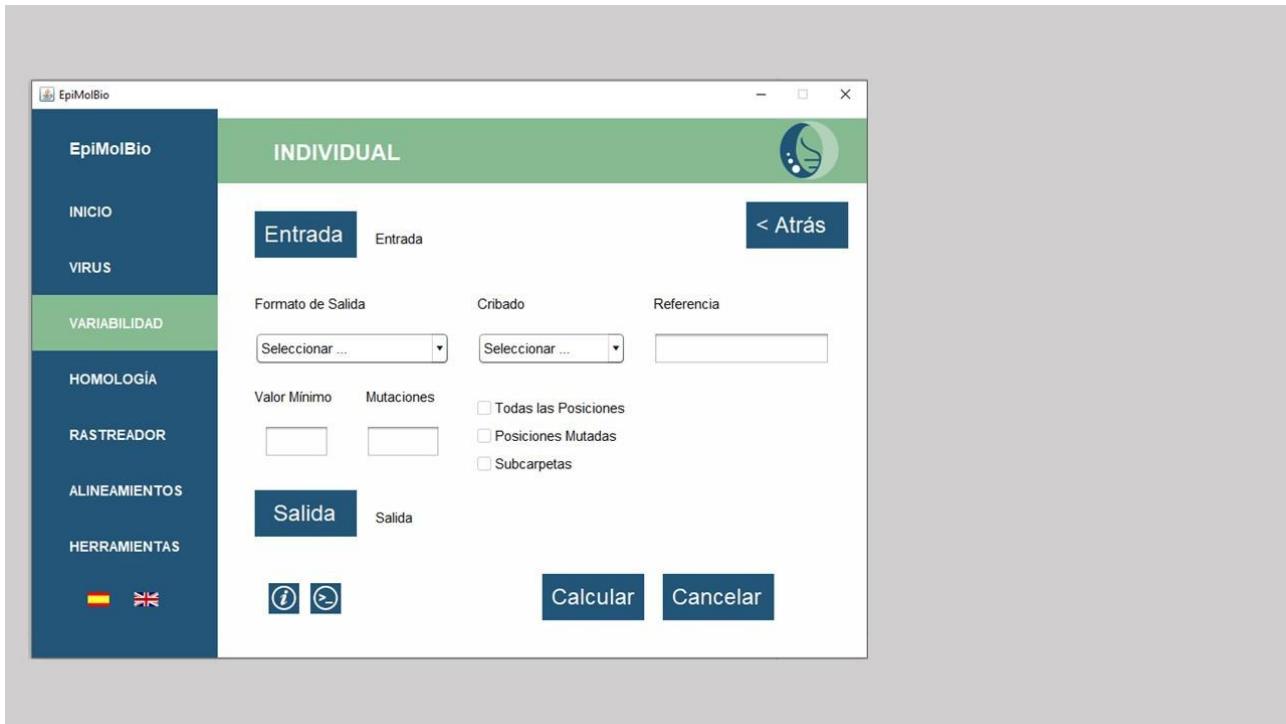
1)



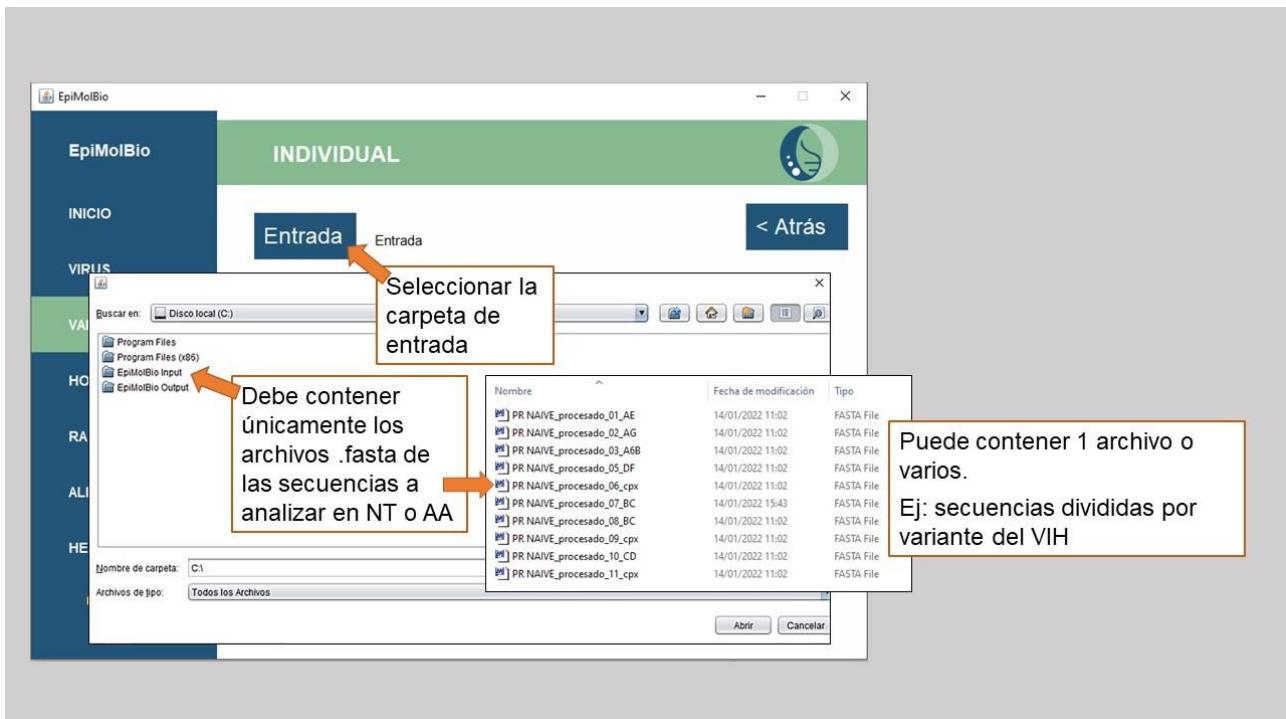
2)



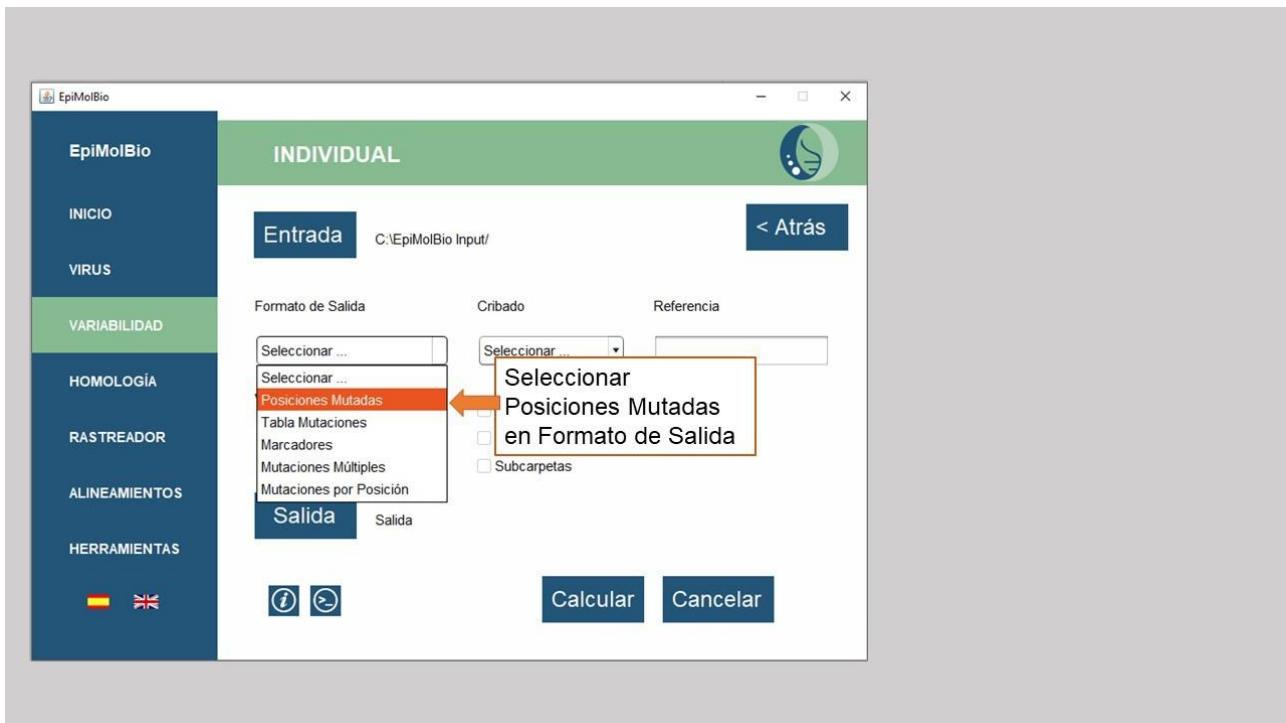
3)



4)



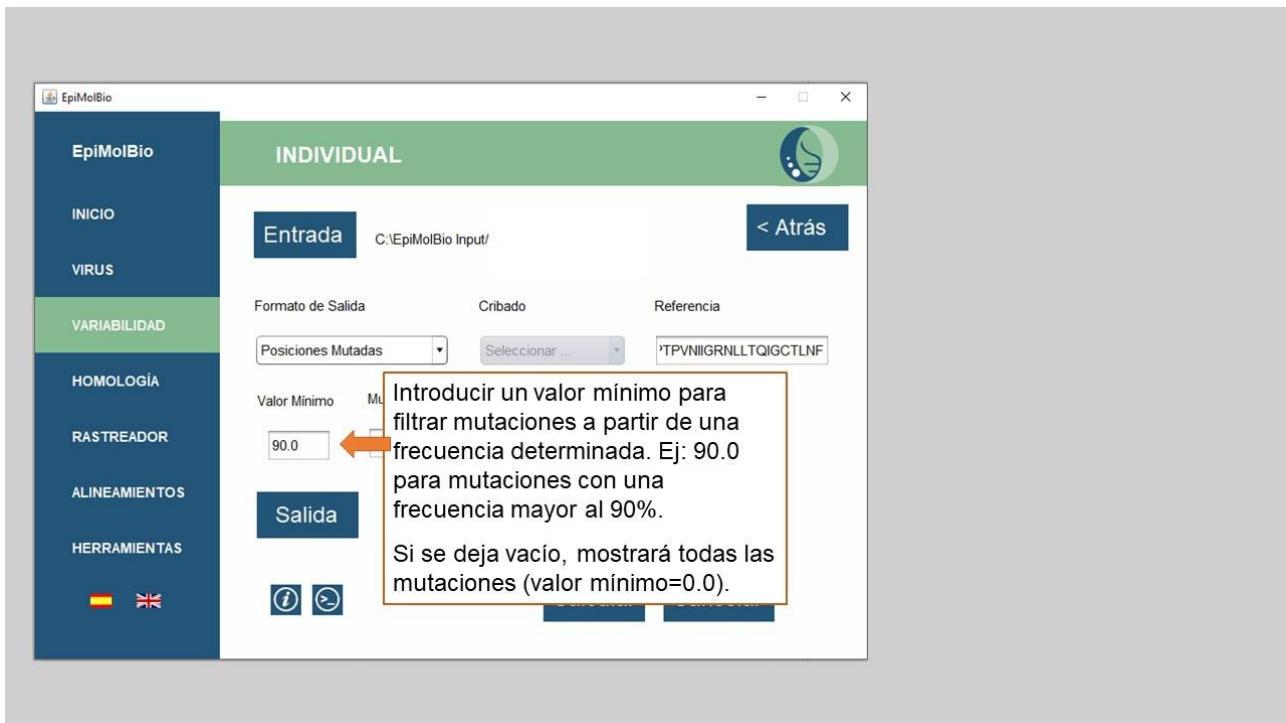
5)



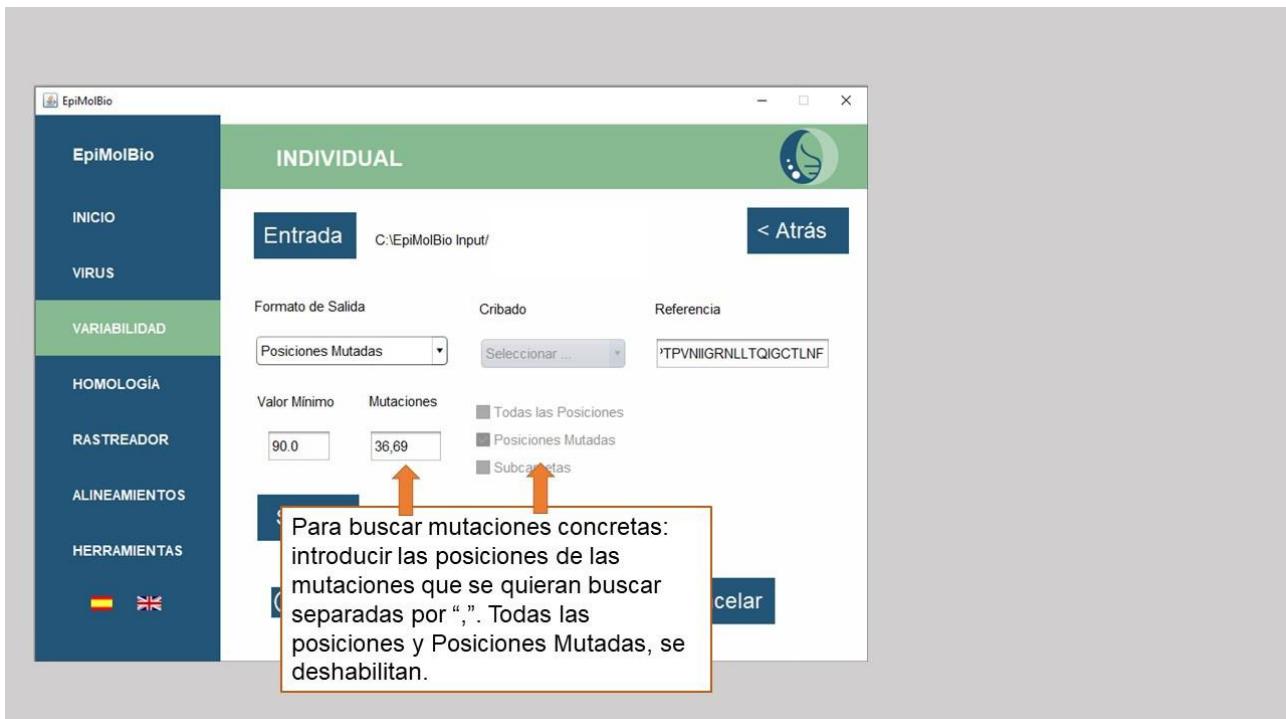
6)



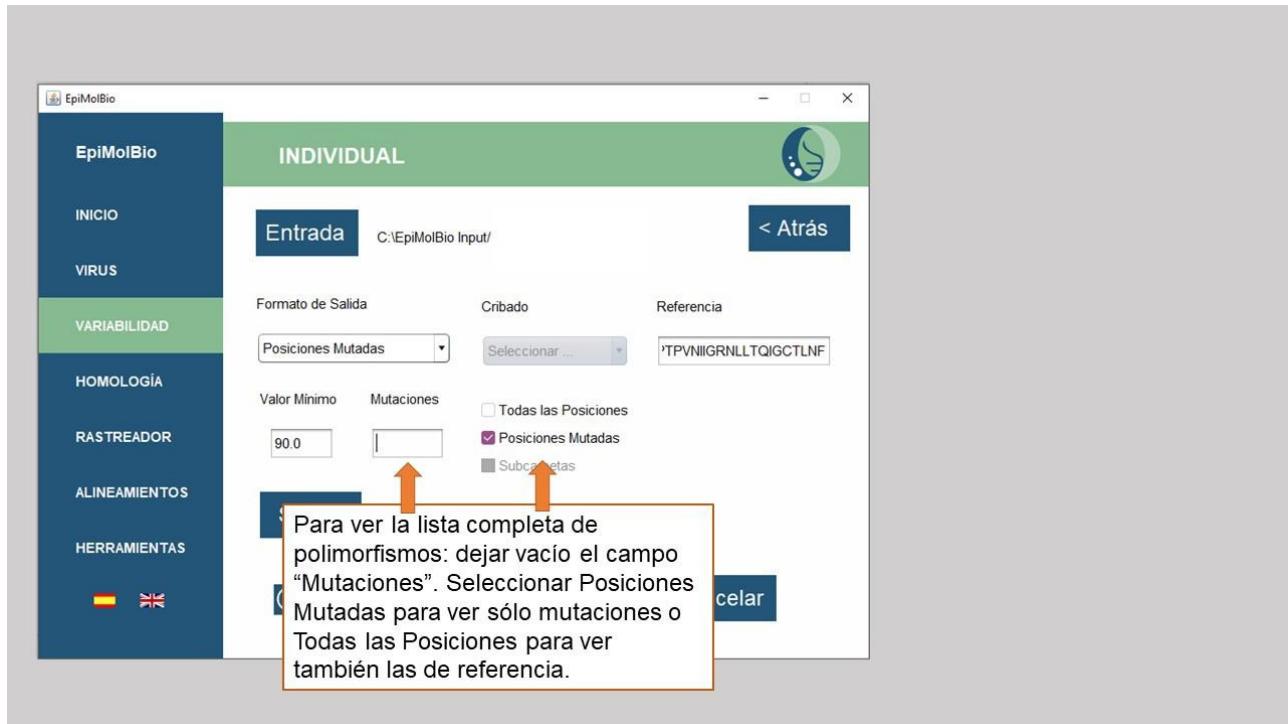
7)



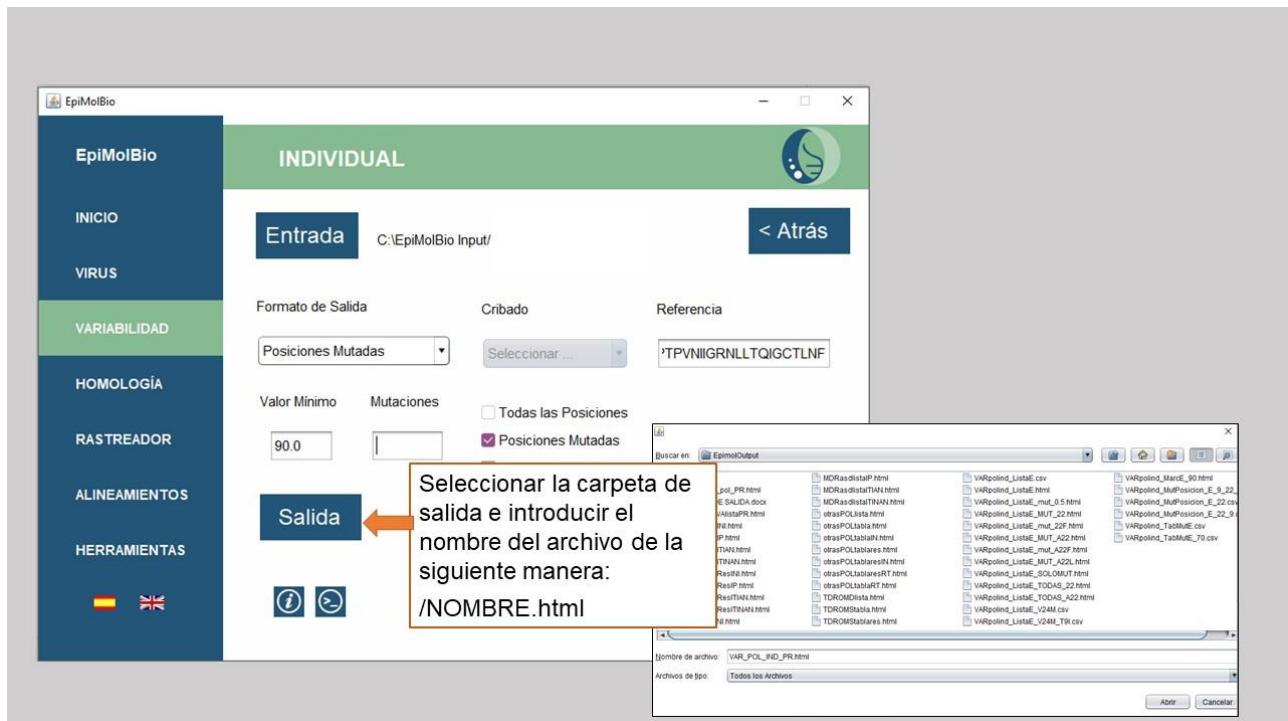
8)



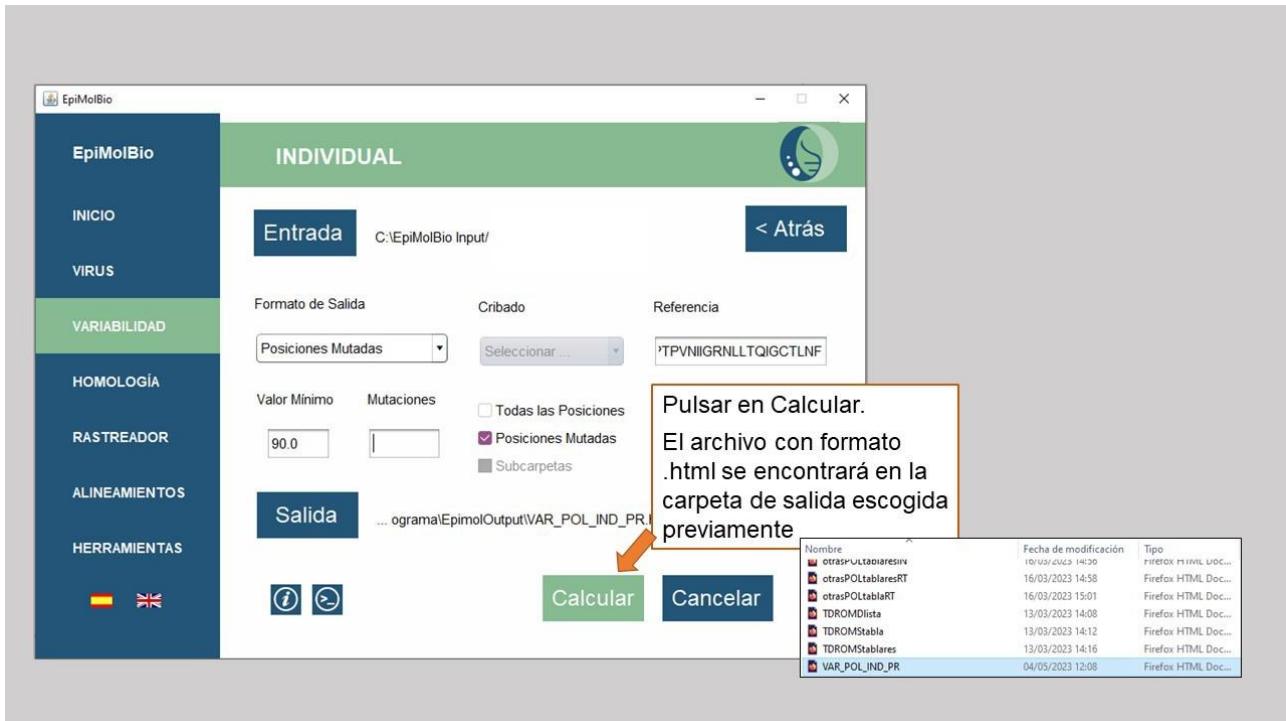
9)



10)



11)



2.-Tabla Mutaciones:

De manera análoga a “Posiciones Mutadas”, con este formato de salida se puede **detectar, localizar y conocer la frecuencia de aparición de mutaciones** con respecto a una secuencia de referencia introducida por el usuario. También se puede escoger buscar sólo una o varias mutaciones concretas, así como seleccionar la frecuencia mínima de aparición de las mutaciones que queremos que se muestren. La diferencia con el formato anterior es que permite introducir los archivos de entrada en subcarpetas y que el archivo de salida es un .csv fácilmente modificable con Excel.

Ejemplo de formato de salida Tabla Mutaciones:

| A | B | C | D | E | F | G | |
|----|----------------|-------------------|------------|----------|--------|------------|--------------------|
| 1 | Archivo | Número Secuencias | Referencia | Posición | Cambio | Porcentaje | Secuencias Mutadas |
| 2 | PR_01_AE.fasta | 26838 | P | 1 | P | 99.90% | 26810 |
| 3 | PR_01_AE.fasta | 26838 | P | 1 | S | 0.08% | 21 |
| 4 | PR_01_AE.fasta | 26838 | P | 1 | A | 0.00% | 1 |
| 5 | PR_01_AE.fasta | 26838 | P | 1 | L | 0.01% | 2 |
| 6 | PR_01_AE.fasta | 26838 | P | 1 | T | 0.01% | 2 |
| 7 | PR_01_AE.fasta | 26838 | P | 1 | H | 0.00% | 1 |
| 8 | PR_01_AE.fasta | 26838 | P | 1 | V | 0.00% | 1 |
| 9 | PR_01_AE.fasta | 26649 | Q | 2 | Q | 99.78% | 26591 |
| 10 | PR_01_AE.fasta | 26649 | Q | 2 | E | 0.07% | 19 |
| 11 | PR_01_AE.fasta | 26649 | Q | 2 | S | 0.02% | 5 |
| 12 | PR_01_AE.fasta | 26649 | Q | 2 | H | 0.06% | 15 |
| 13 | PR_01_AE.fasta | 26649 | Q | 2 | D | 0.00% | 1 |
| 14 | PR_01_AE.fasta | 26649 | Q | 2 | K | 0.02% | 6 |
| 15 | PR_01_AE.fasta | 26649 | Q | 2 | L | 0.02% | 4 |
| 16 | PR_01_AE.fasta | 26649 | Q | 2 | P | 0.01% | 2 |
| 17 | PR_01_AE.fasta | 26649 | Q | 2 | R | 0.01% | 3 |
| 18 | PR_01_AE.fasta | 26649 | Q | 2 | T | 0.00% | 1 |
| 19 | PR_01_AE.fasta | 26649 | Q | 2 | * | 0.01% | 2 |
| 20 | PR_01_AE.fasta | 26831 | V | 3 | I | 99.86% | 26793 |

El archivo .csv de salida es una tabla que contiene, en la primera columna aparece el nombre de los archivos de entrada; en la segunda, el número de secuencias totales para cada archivo; en la tercera, el aminoácido o nucleótido de referencia según la secuencia de referencia introducida; en la siguiente columna, las posiciones donde se han detectado mutaciones; a continuación, el cambio de aminoácido o nucleótido detectado; en la sexta columna la frecuencia de aparición del cambio detectado y, en la última columna, el número total de secuencias mutadas. Al tratarse de un archivo .csv, puede manipularse con Excel para aplicar filtros o unir columnas si se desea. Si se escoge el campo “Todas las Posiciones”, aparecerán todas aunque no estén mutadas y también aparecerá el porcentaje del aminoácido de referencia.

Para realizar este análisis, en **entrada** se debe seleccionar una carpeta donde se tengan exclusivamente los archivos en formato .fasta o una carpeta que contenga otras subcarpetas con archivos .fasta. Para esto se debe marcar la opción de Subcarpetas. Los archivos de entrada pueden estar en nucleótidos o aminoácidos. Para realizar el análisis en secuencias de nucleótidos será necesario emplear la herramienta “Buscar y Reemplazar” de “Edición de Archivos” en “Herramientas” y sustituir las “N” por “?” para que estas se excluyan del análisis.

Habrá que seleccionar “**Tabla de mutaciones**” en el desplegable “**Formato de salida**”.

En “**Referencia**” introducir la secuencia de referencia en letras sin saltos de línea, teniendo en cuenta que la secuencia debe introducirse en nucleótidos o aminoácidos según si los archivos de entrada están sin traducir o traducidos.

El campo “**Valor Mínimo**” puede dejarse vacío si se desea que el resultado muestre todas las mutaciones detectadas, independientemente de su frecuencia de aparición. Si se quiere filtrar sólo mutaciones que aparezcan en cierta frecuencia, habrá que introducir el valor mínimo en formato numérico decimal (ej.: 75.0 para que aparezcan sólo las mutaciones que se encuentran en un porcentaje mayor al 75%).

El campo “**Mutaciones**” puede dejarse vacío si se desea que el resultado muestre todas las mutaciones detectadas. Si se quiere buscar una o varias mutaciones concretas, habrá que introducir la mutación tecleando la referencia, posición y residuo mutado (ej.: Q2H). Cuando los archivos de entrada y la referencia estén en nucleótidos, introducir la mutación con la referencia y la mutación en nucleótidos (ej.: A6C). Si se quieren buscar varias mutaciones, separarlas por una coma “,” sin espacios.

Existe la posibilidad de ver lo que hay en todas las posiciones o sólo en las posiciones mutadas. Para ello se debe seleccionar la casilla “**Todas las Posiciones**” o “**Posiciones Mutadas**” según corresponda. Si se ha introducido algo en el campo “**Mutaciones**”, estas cajas se deshabilitan.

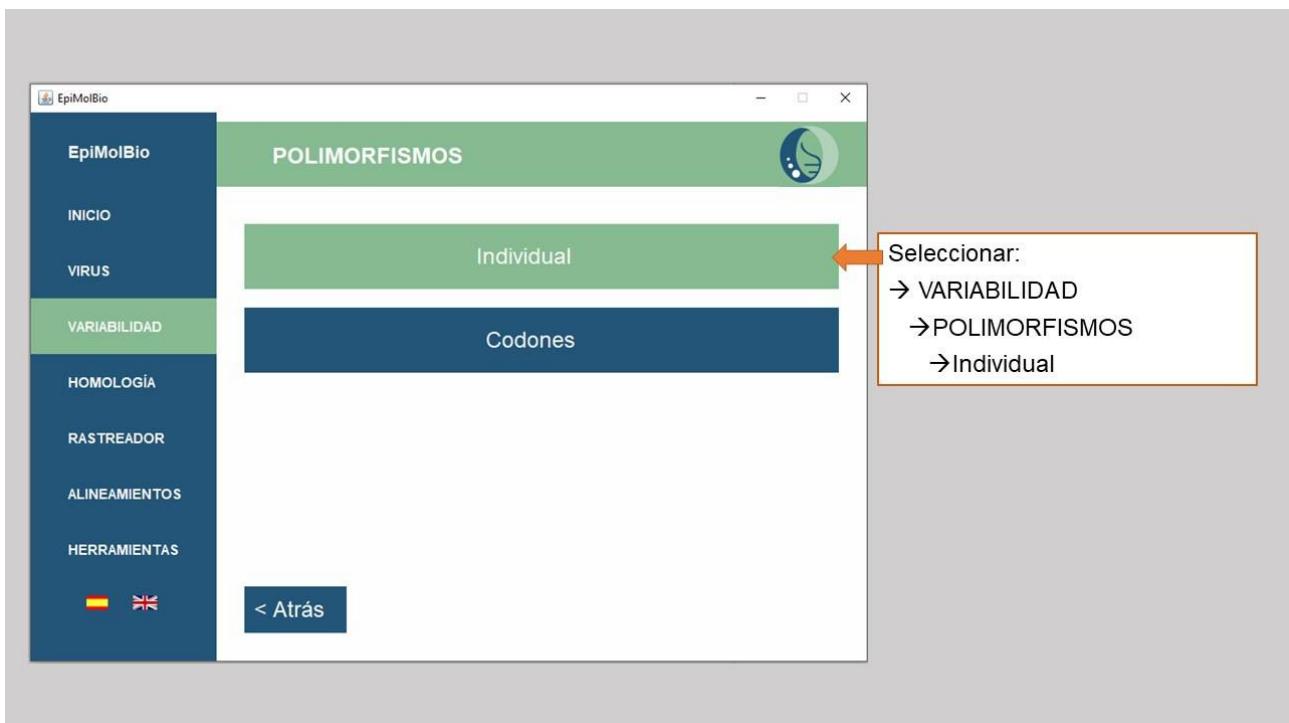
El resultado se muestra en un archivo .csv. En **salida** se debe seleccionar la carpeta donde se quiera guardar el resultado y nombrar el archivo con la extensión .csv.

Paso a paso:

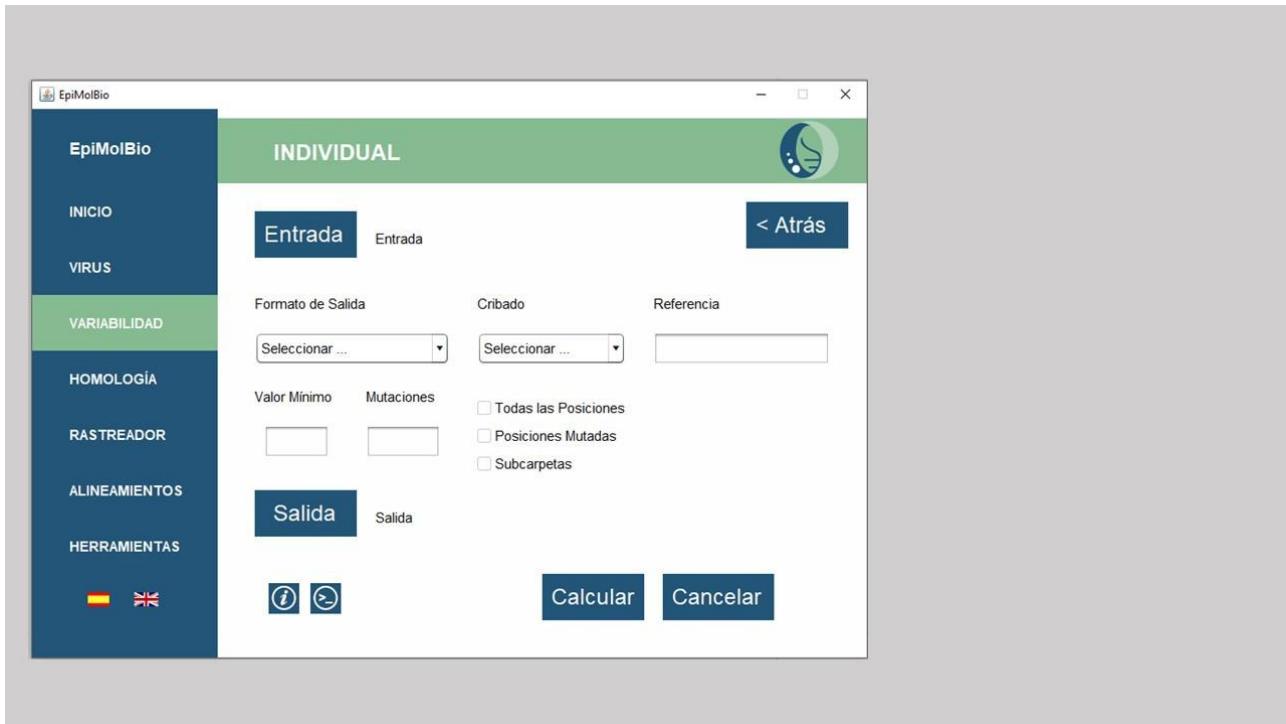
1)



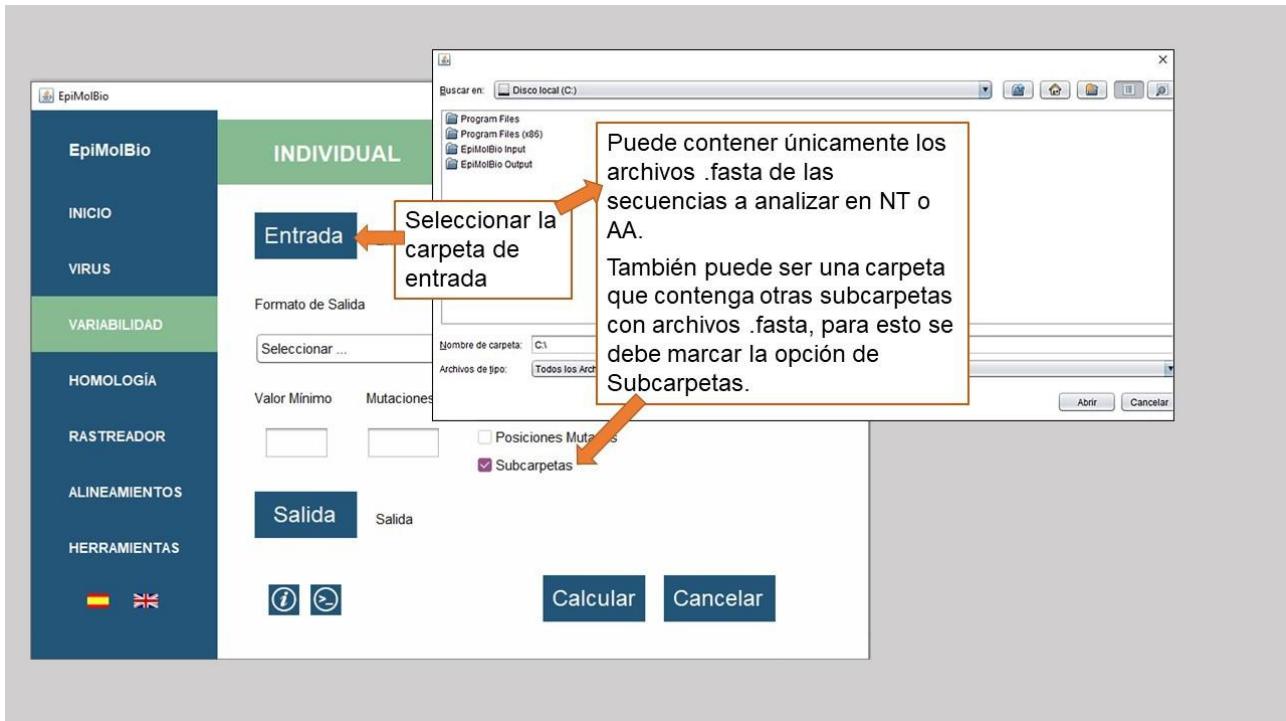
2)



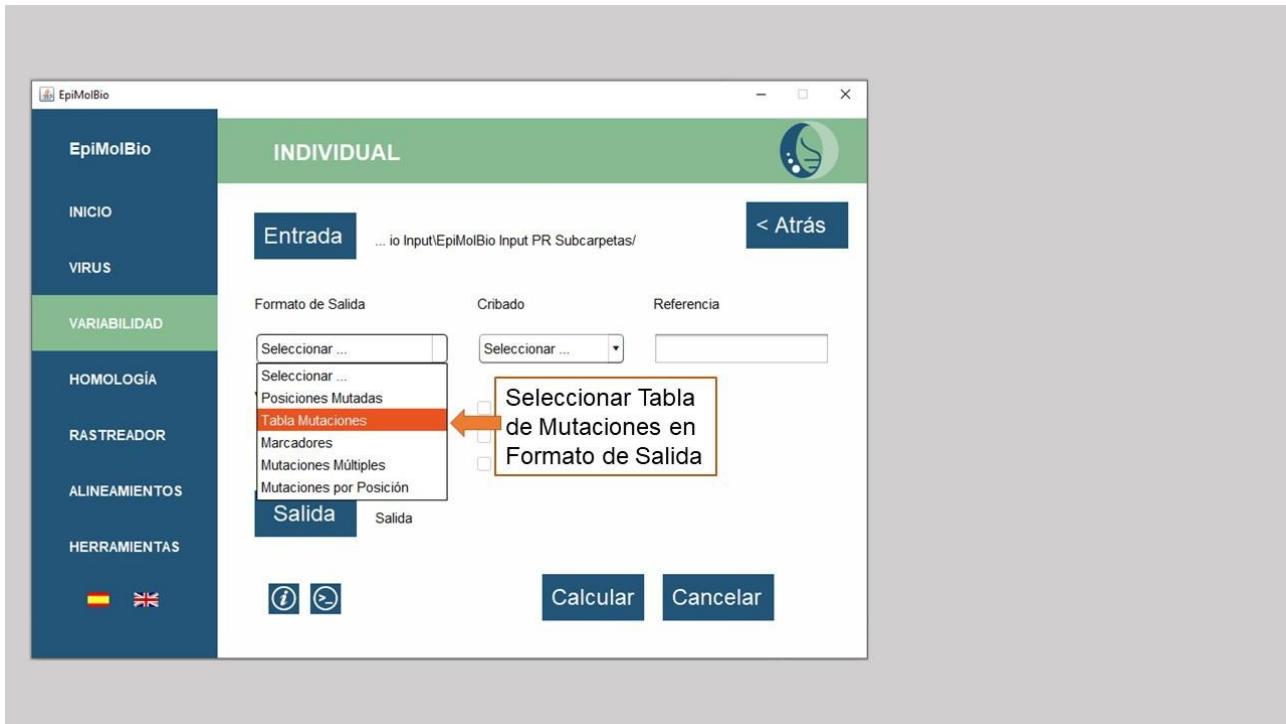
3)



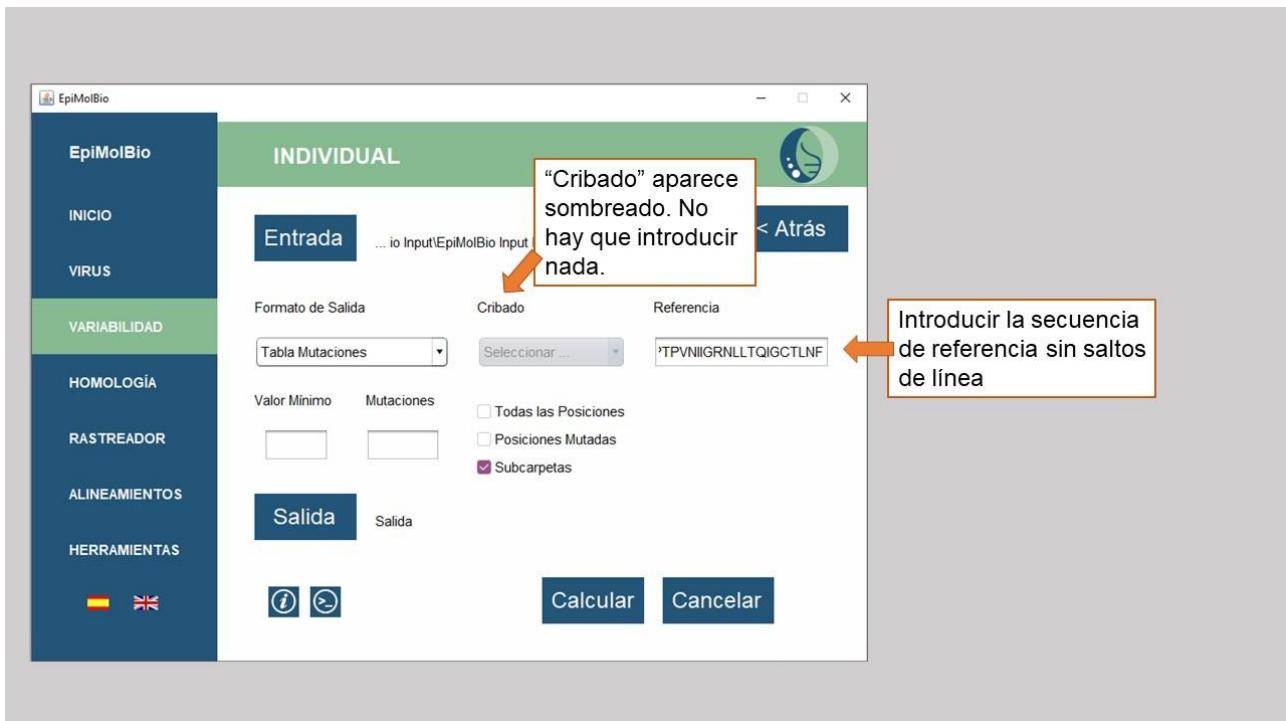
4)



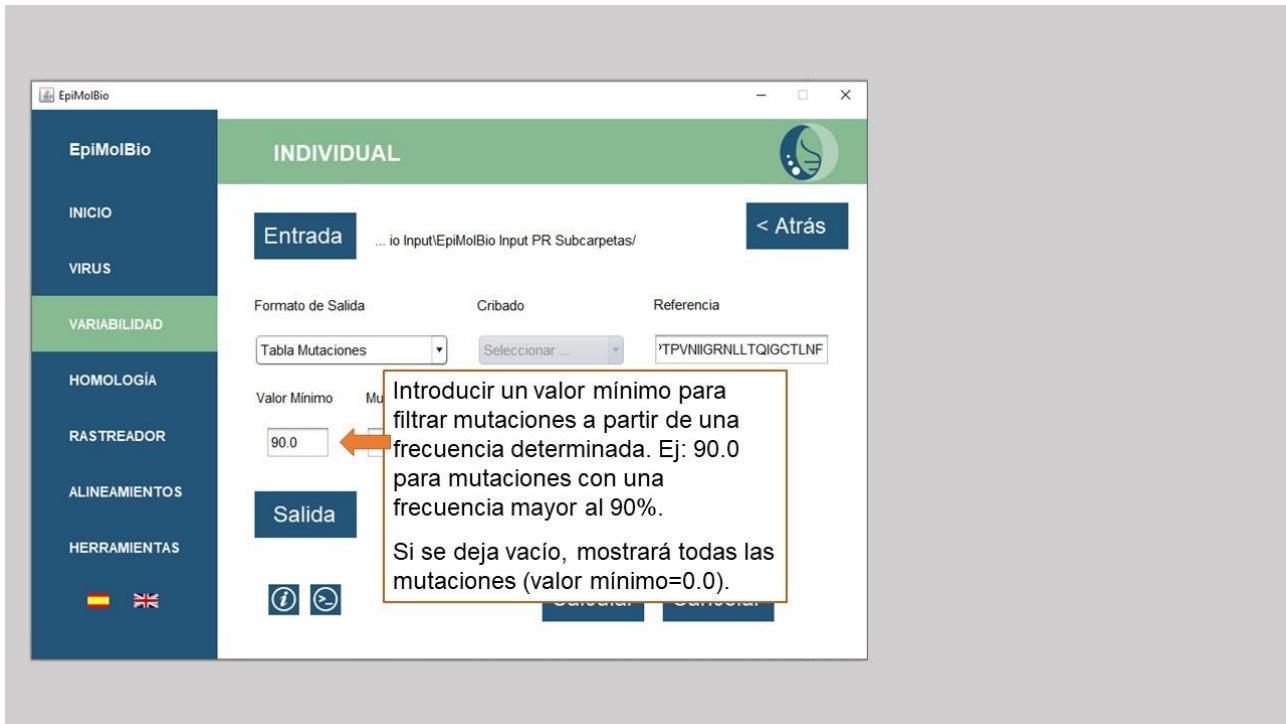
5)



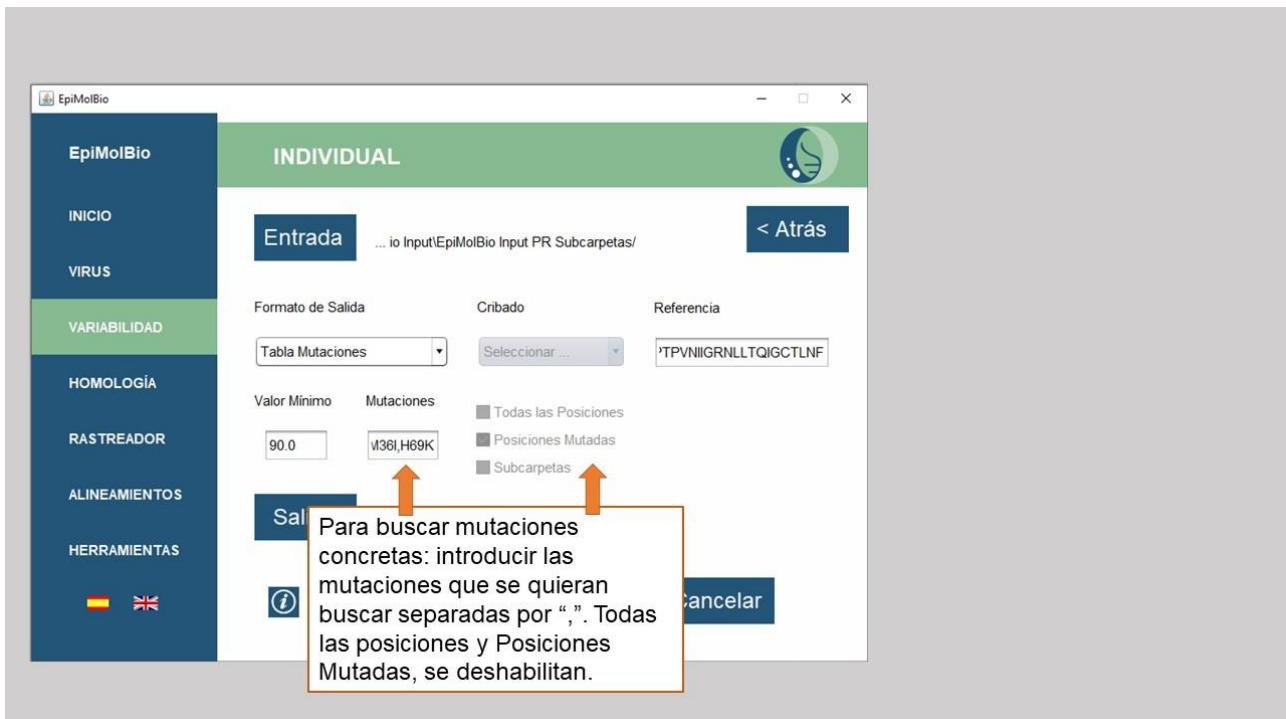
6)



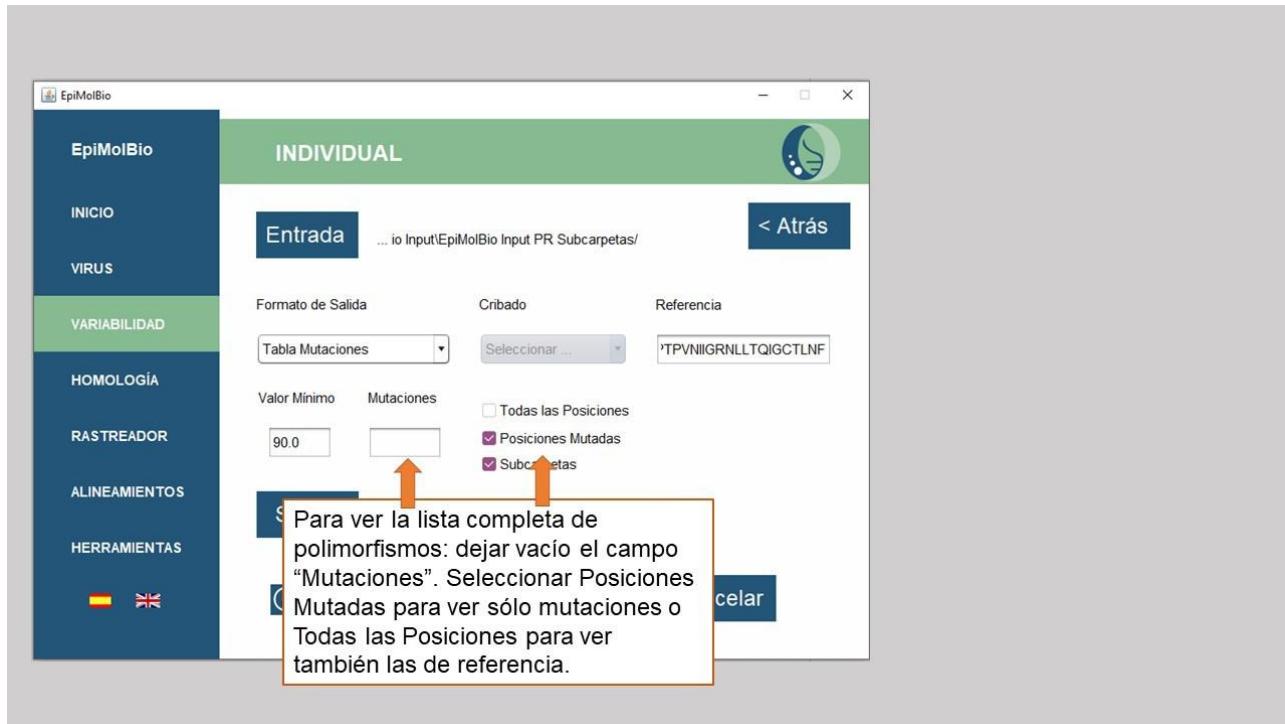
7)



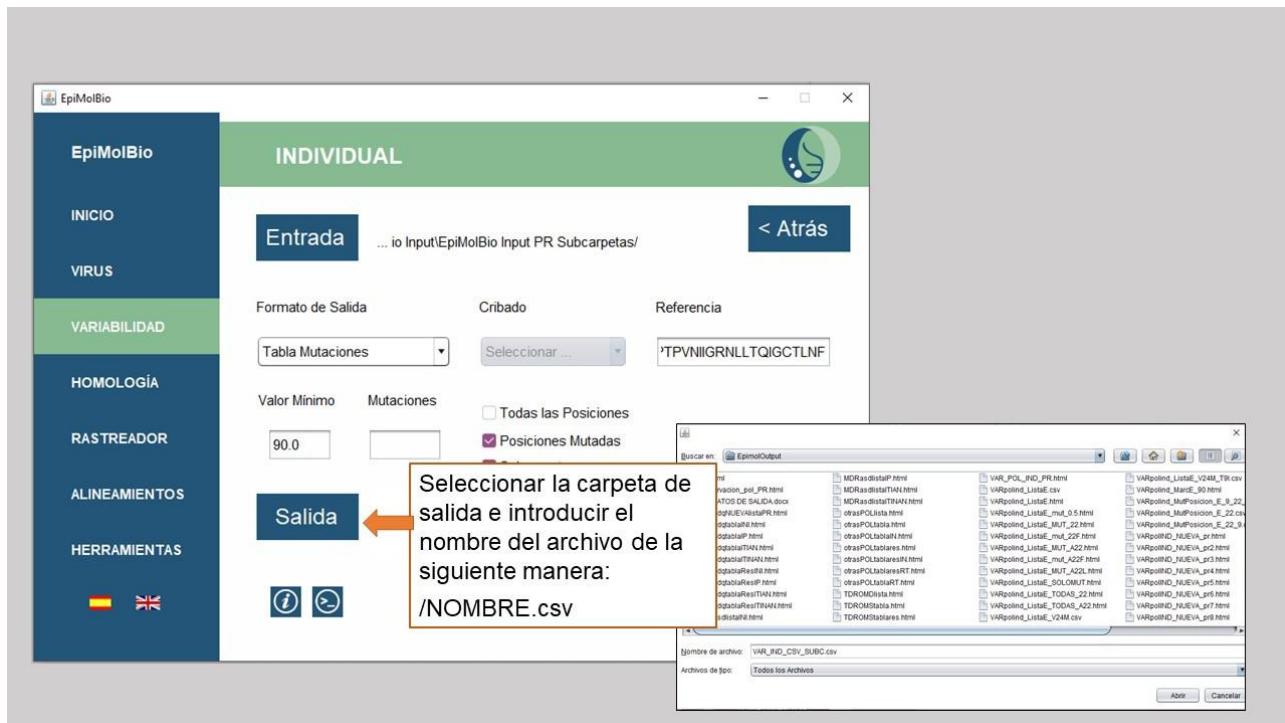
8)



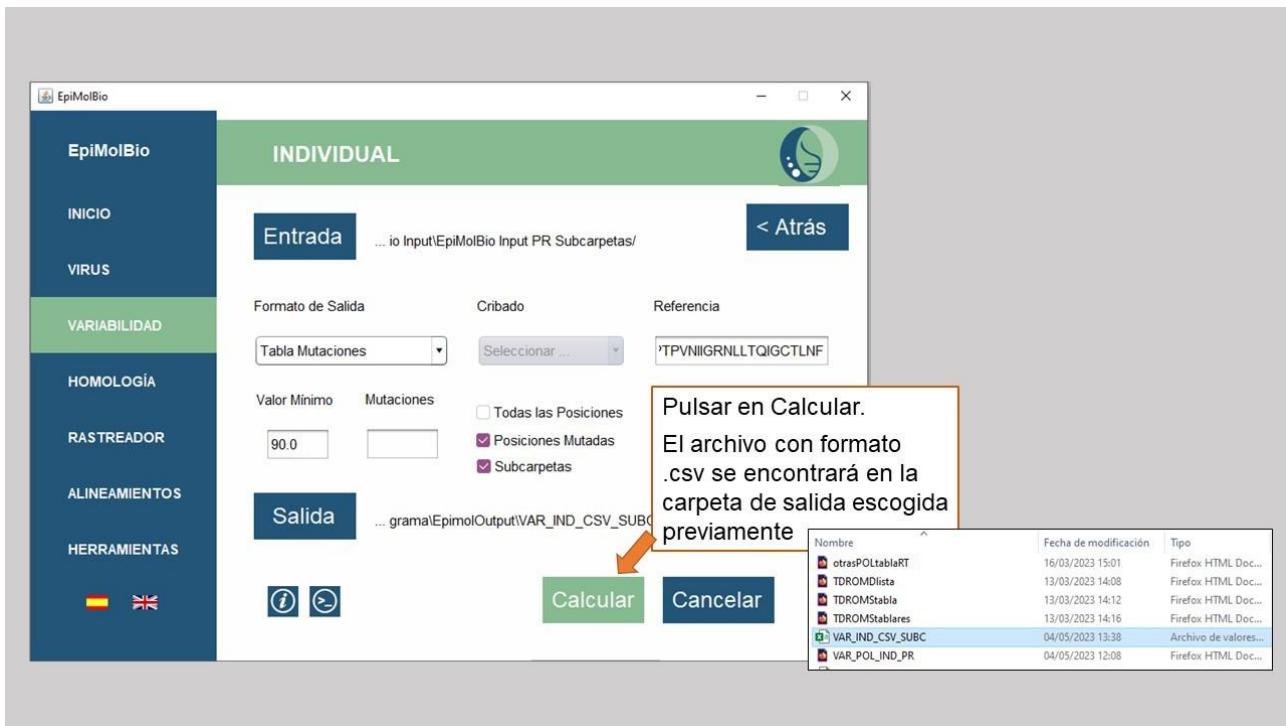
9)



10)



11)



3.- Marcadores:

Permite **detectar las mutaciones exclusivas** de cada archivo en comparación al resto de archivos introducidos como entrada. Se puede establecer si se quiere que las mutaciones estén presentes en una frecuencia superior al 75% o el 90%. Si, por ejemplo, los archivos están divididos por variantes virales, podremos encontrar mutaciones características de cada variante con esta función. Estas mutaciones características son a las que nos referiremos como marcadores. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis.

En el archivo de salida aparece, en la parte superior, el título del análisis. Debajo, en la columna “Archivo”, aparece el nombre del archivo de entrada analizado, en la segunda, “Marcadores”, se muestran los marcadores que se hayan detectado según el porcentaje de cribado seleccionado y en la tercera, “Secuencias Totales”, aparece el número todas de secuencias analizadas. Si se encuentran marcadores para el archivo, se indicará el aminoácido de referencia, seguido de la posición y la mutación coloreada según el código de colores descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

Ejemplo de formato de salida Marcadores:

| Variabilidad Polimorfismos Individual Marcadores >= 90% | | |
|---|------------|--------------------|
| Archivo | Marcadores | Secuencias Totales |
| PR_107_01B.fasta | Q92K | 4 |
| PR_108_BC.fasta | T74S | 15 |
| PR_112_01B.fasta | L63M | 5 |
| PR_118_BC.fasta | K70T | 13 |
| PR_11_cpx.fasta | G16A | 380 |
| PR_129_56G.fasta | K14D, E65K | 1 |

Para realizar este análisis, en **entrada** se debe seleccionar una **carpeta** donde se tengan exclusivamente los archivos en formato .fasta de las secuencias a analizar en **aminoácidos**.

Habrá que seleccionar “**Marcadores**” en el desplegable “**Formato de salida**”.

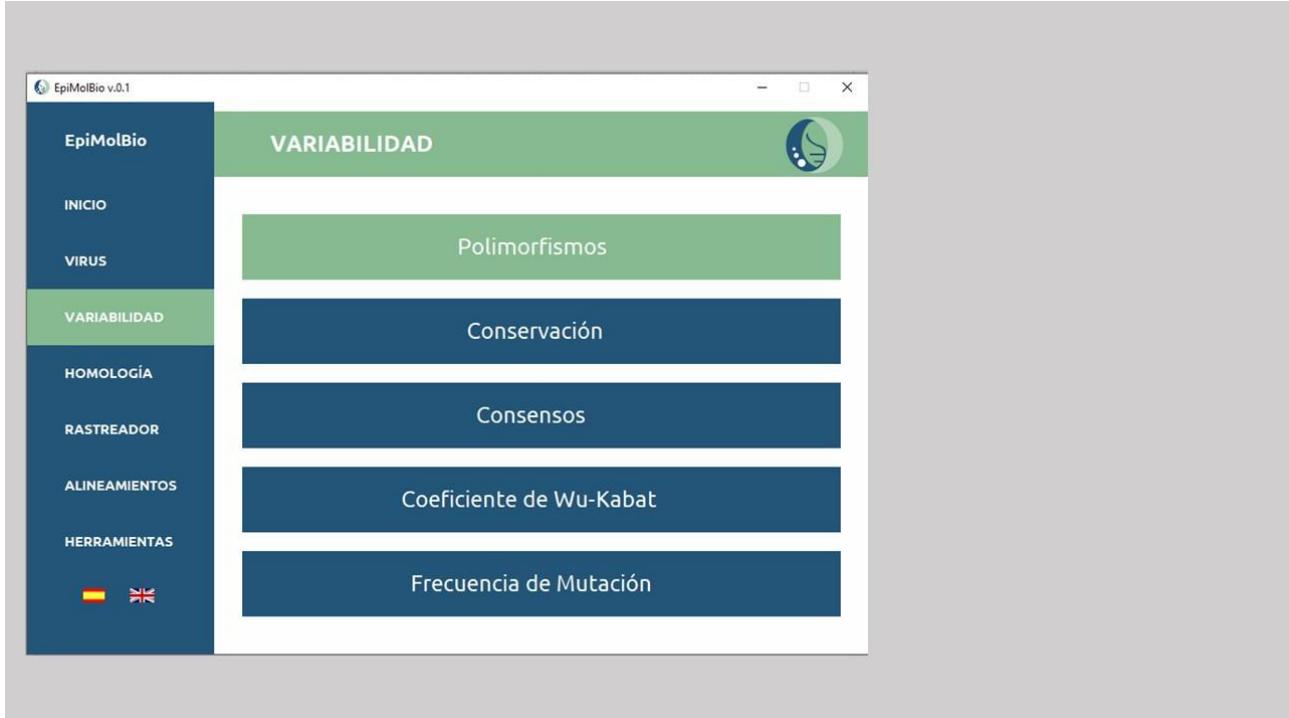
En el campo “**Cribado**” seleccionar si se quiere que los marcadores detectados estén en una frecuencia superior al 75% (Mostrar > 75%) o el 90% (Mostrar >= 90%).

En “**Referencia**” introducir la secuencia de referencia en letras sin saltos de línea.

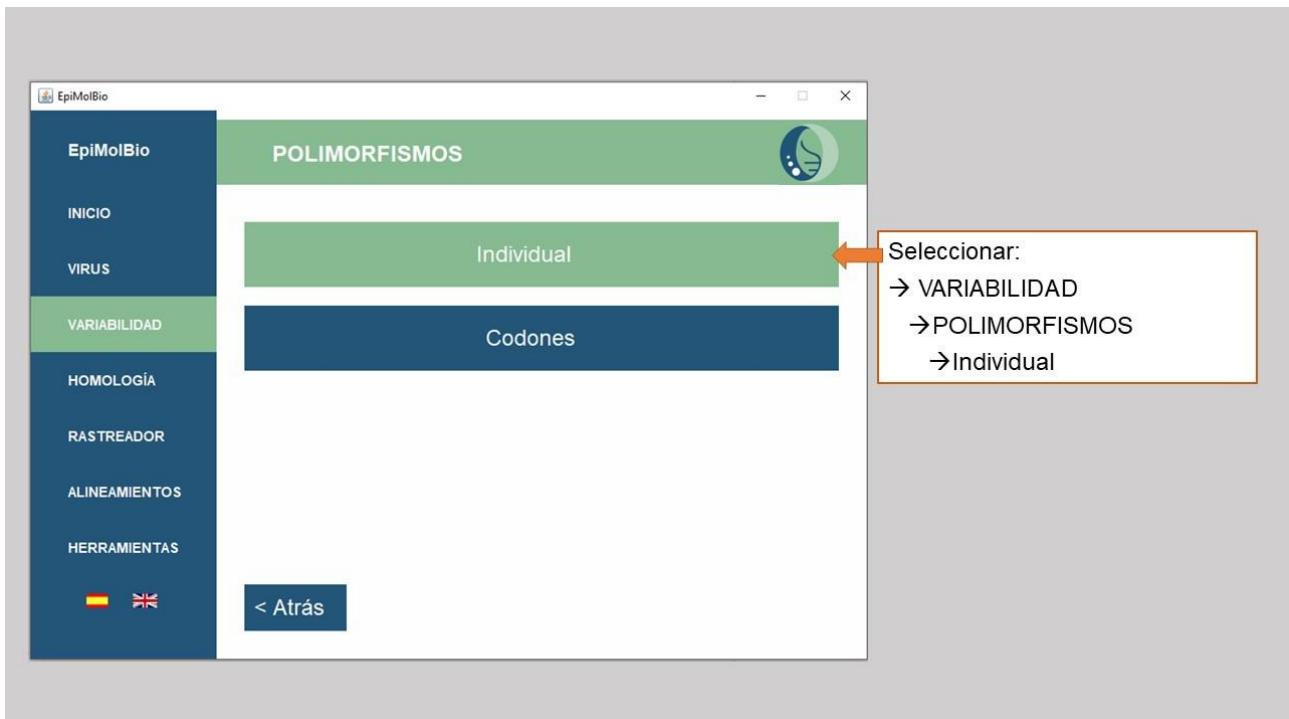
El resultado se muestra en un archivo .html. En **salida** se debe seleccionar la carpeta donde se quiera guardar el resultado y nombrar el archivo con la extensión .html.

Paso a paso:

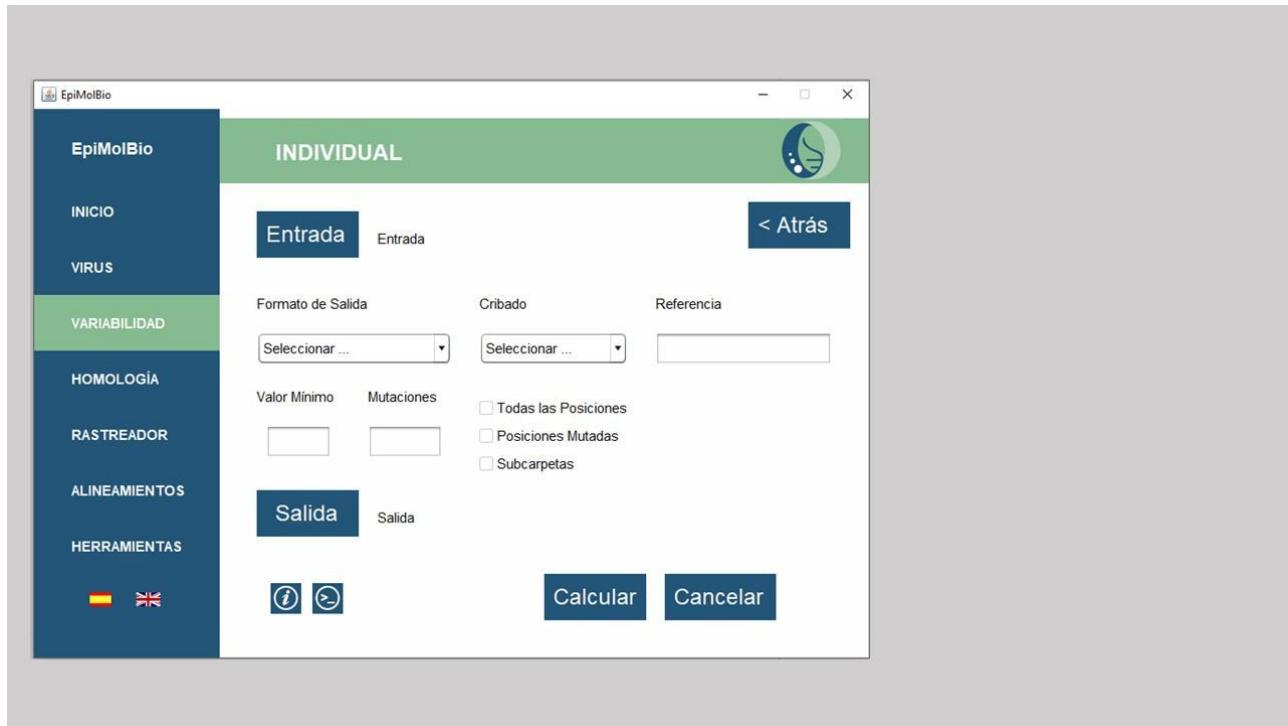
1)



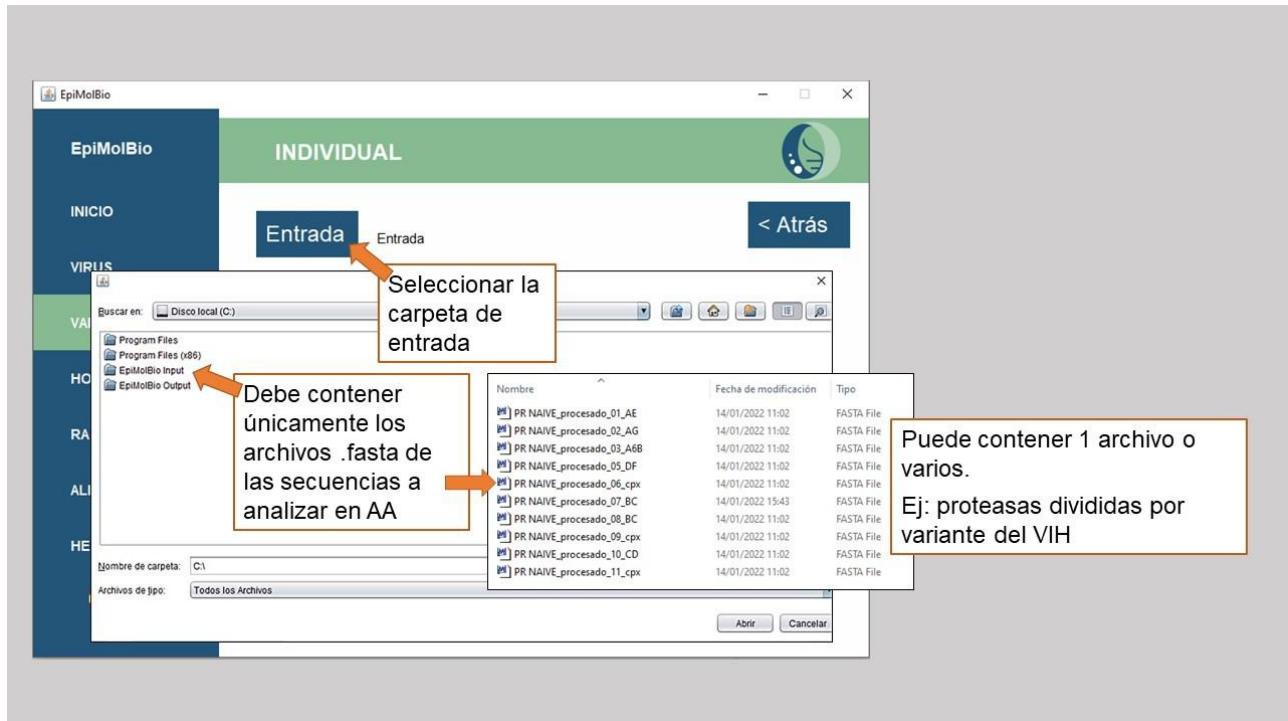
2)



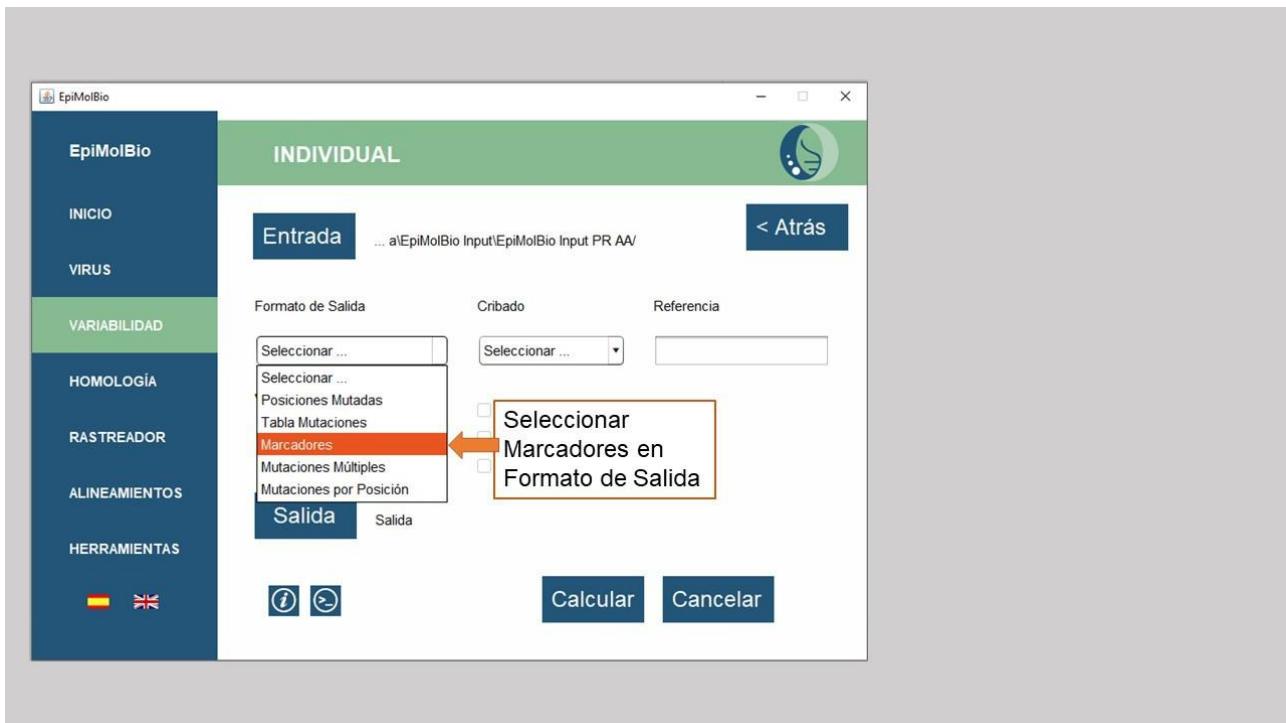
3)



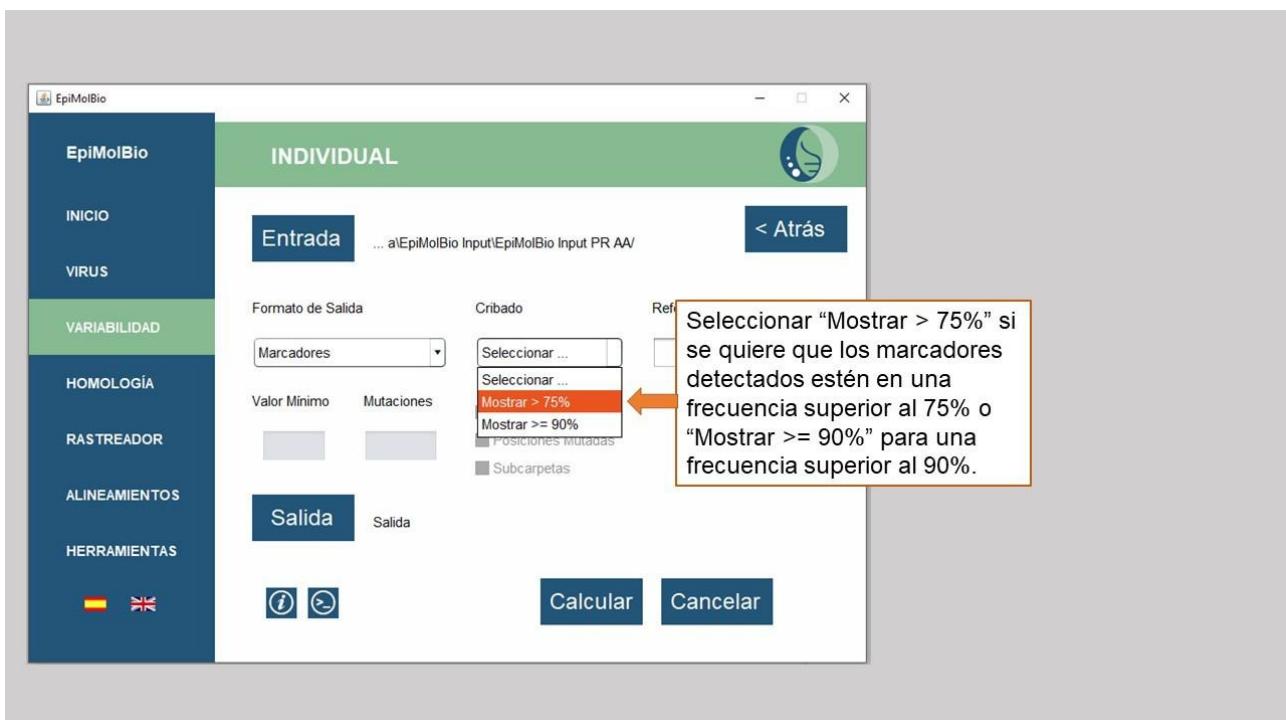
4)



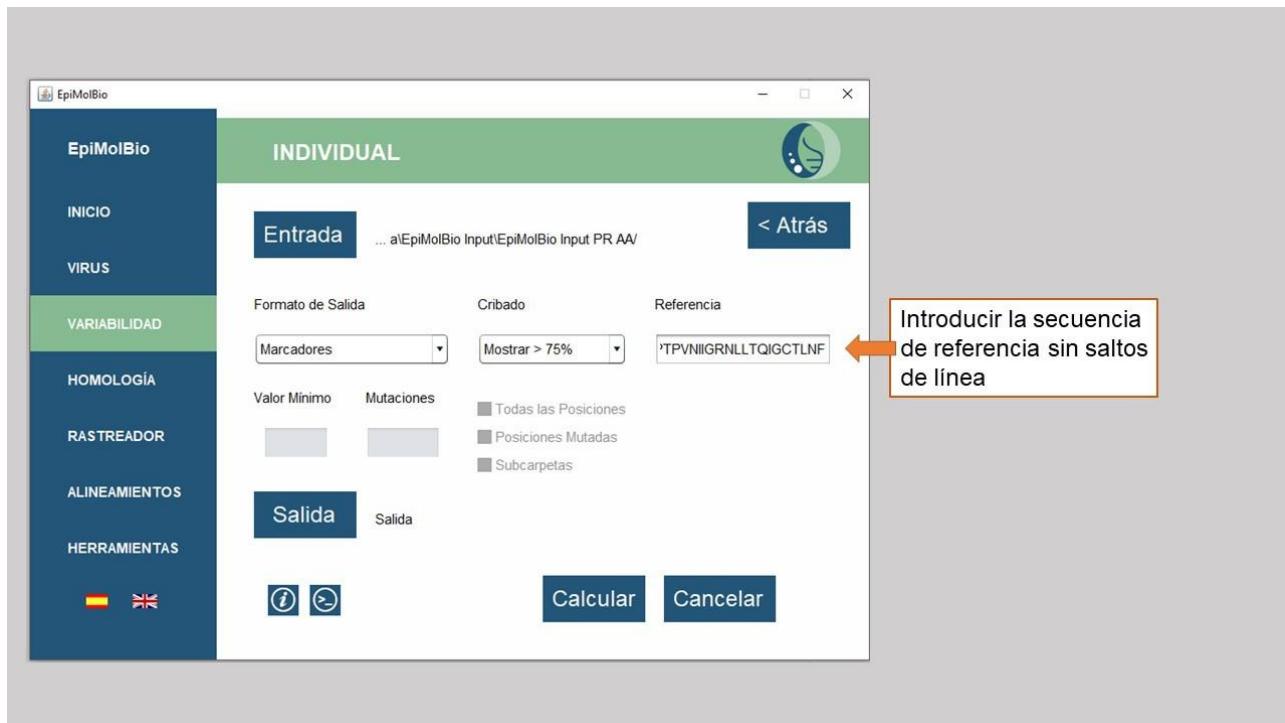
5)



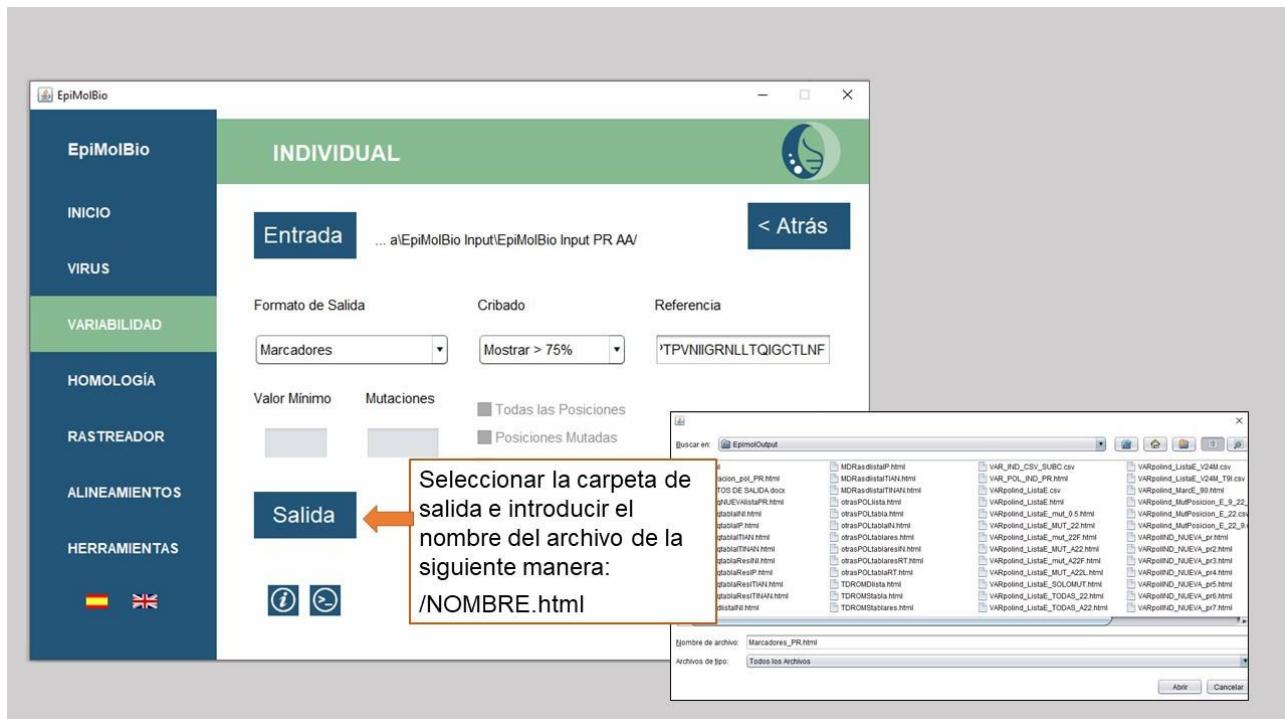
6)



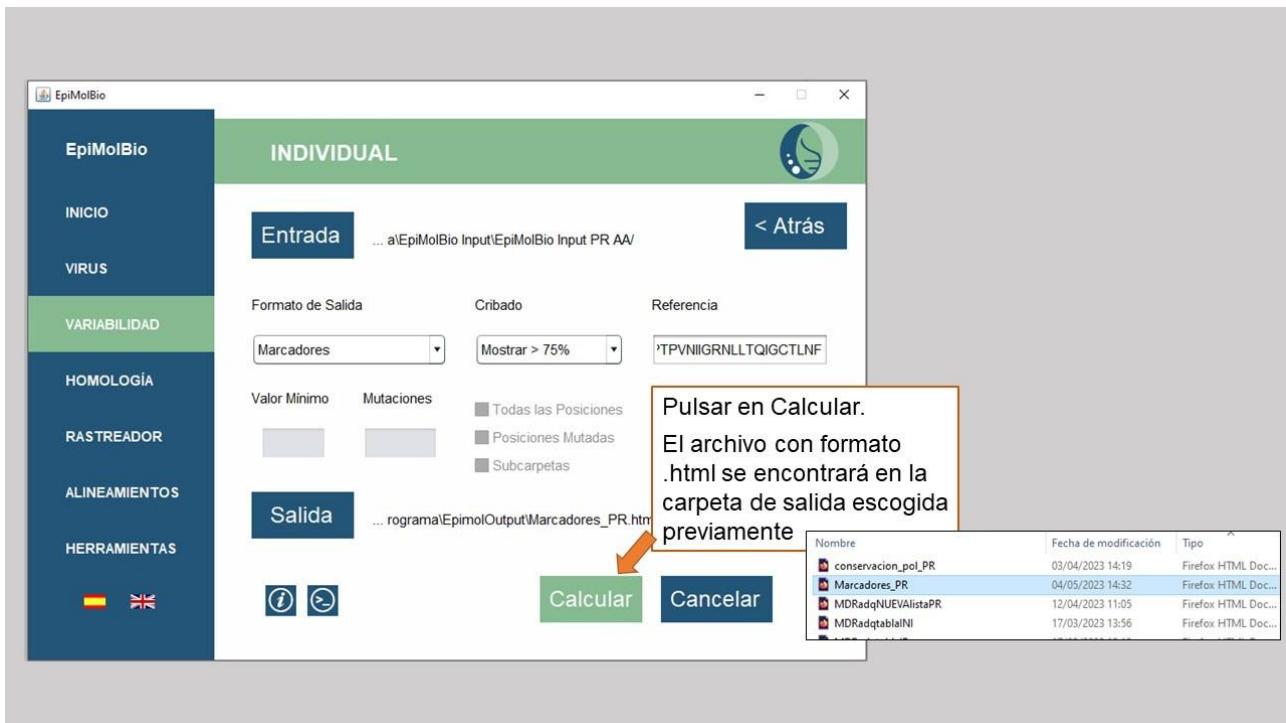
7)



8)



9)



4.-Mutaciones Múltiples:

Permite **detectar y conocer la frecuencia de aparición de combinaciones de mutaciones** que deben introducirse en el campo “Mutaciones”. A su vez, permite introducir los archivos de entrada desde subcarpetas.

El archivo .csv de salida es una tabla que contiene, en la primera columna, el nombre de los archivos de entrada; en la segunda, el número total de secuencias válidas por archivo; en la tercera, el número de veces que aparecen las mutaciones combinadas y en la cuarta la frecuencia de aparición de esa combinación de mutaciones. Sirve tanto para combinaciones de aminoácidos como de nucleótidos. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis.

Ejemplo de formato de salida Mutaciones Múltiples:

| A | B | | C | D |
|----|------------------|-------------------|-----------|------------|
| 1 | Archivo | Número Secuencias | M46I/V32I | Frecuencia |
| 2 | PR_01_AE.fasta | 26504 | 2 | 0.008 |
| 3 | PR_02_AG.fasta | 9418 | 0 | 0 |
| 4 | PR_03_A6B.fasta | 300 | 1 | 0.333 |
| 5 | PR_04_cpx.fasta | 15 | 0 | 0 |
| 6 | PR_05_DF.fasta | 24 | 0 | 0 |
| 7 | PR_06_cpx.fasta | 732 | 0 | 0 |
| 8 | PR_07_BC.fasta | 10819 | 0 | 0 |
| 9 | PR_08_BC.fasta | 2326 | 0 | 0 |
| 10 | PR_09_cpx.fasta | 91 | 0 | 0 |
| 11 | PR_100_01C.fasta | 5 | 0 | 0 |
| 12 | PR_101_01B.fasta | 4 | 0 | 0 |

Para realizar este análisis, en **entrada** se debe seleccionar una carpeta donde se tengan exclusivamente los archivos en formato .fasta en aminoácidos o nucleótidos o una carpeta que contenga otras subcarpetas con archivos .fasta. Para esto se debe marcar la opción de Subcarpetas. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis. Si se quiere realizar el análisis en nucleótidos, se puede emplear la función “Buscar y Reemplazar” de “Herramientas”, “Edición de Archivos”, para cambiar las “N” por “?”, ya que la función Mutaciones Múltiples no excluye las “N” del análisis.

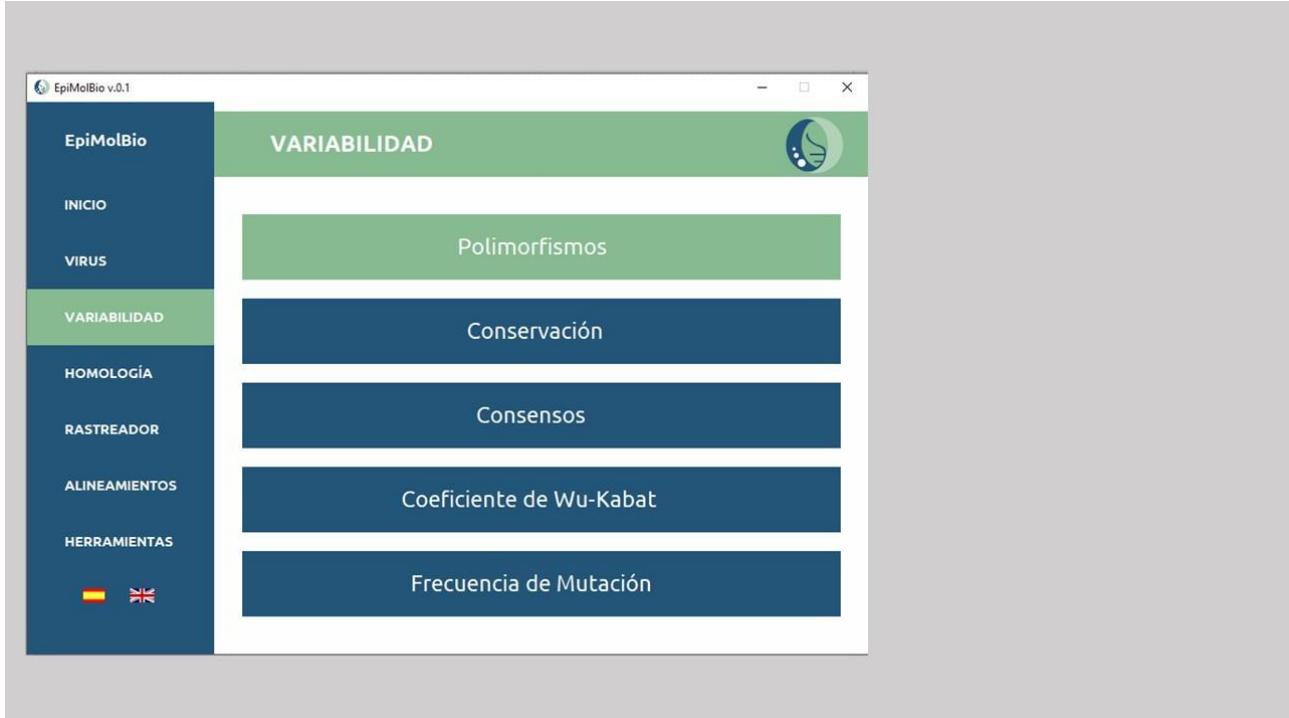
Habrá que seleccionar “**Mutaciones Múltiples**” en el desplegable “**Formato de salida**”.

En el campo “**Mutaciones**”, introducir la combinación de mutaciones que se desea buscar tecleando la referencia, posición y residuo mutado de cada mutación separadas por una coma “,” sin espacios (ej.: D614G,A222V). Cuando los archivos de entrada y la referencia estén en nucleótidos, introducir la mutación con la referencia y la mutación en nucleótidos.

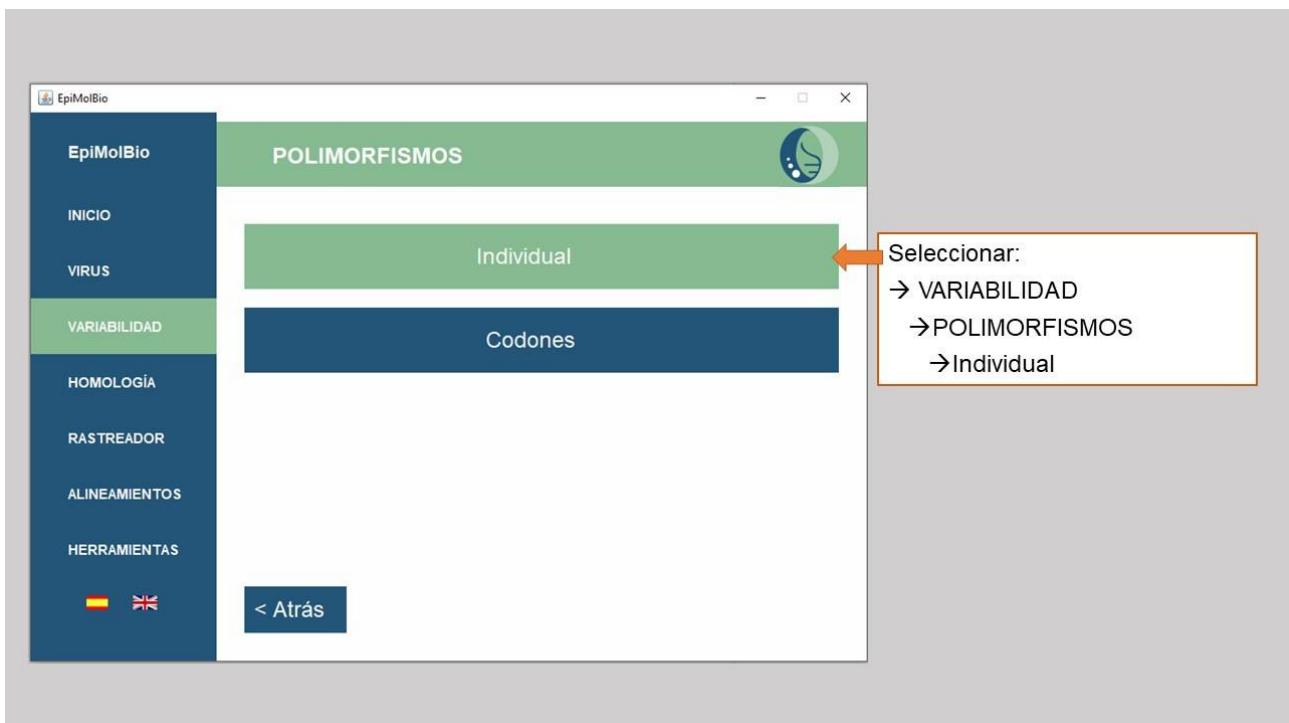
El resultado se muestra en un archivo .csv. En **salida** se debe seleccionar la carpeta donde se quiera guardar el resultado y nombrar el archivo con la extensión .csv.

Paso a paso:

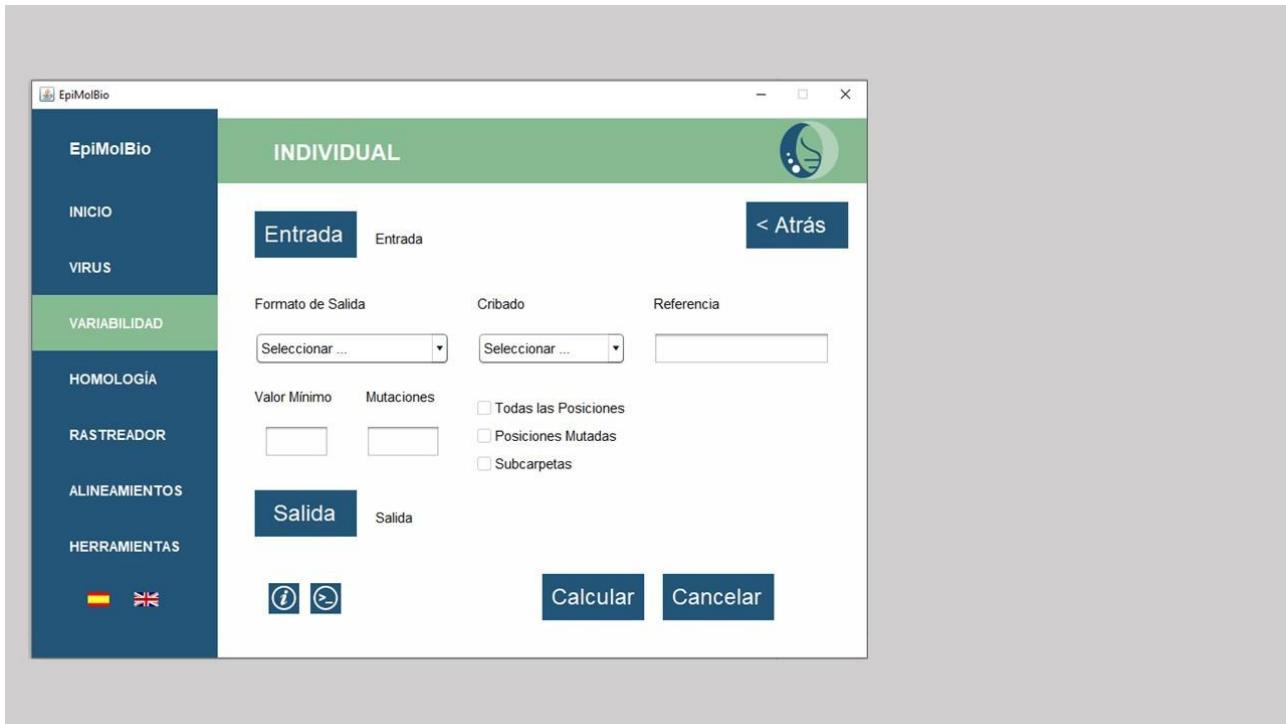
1)



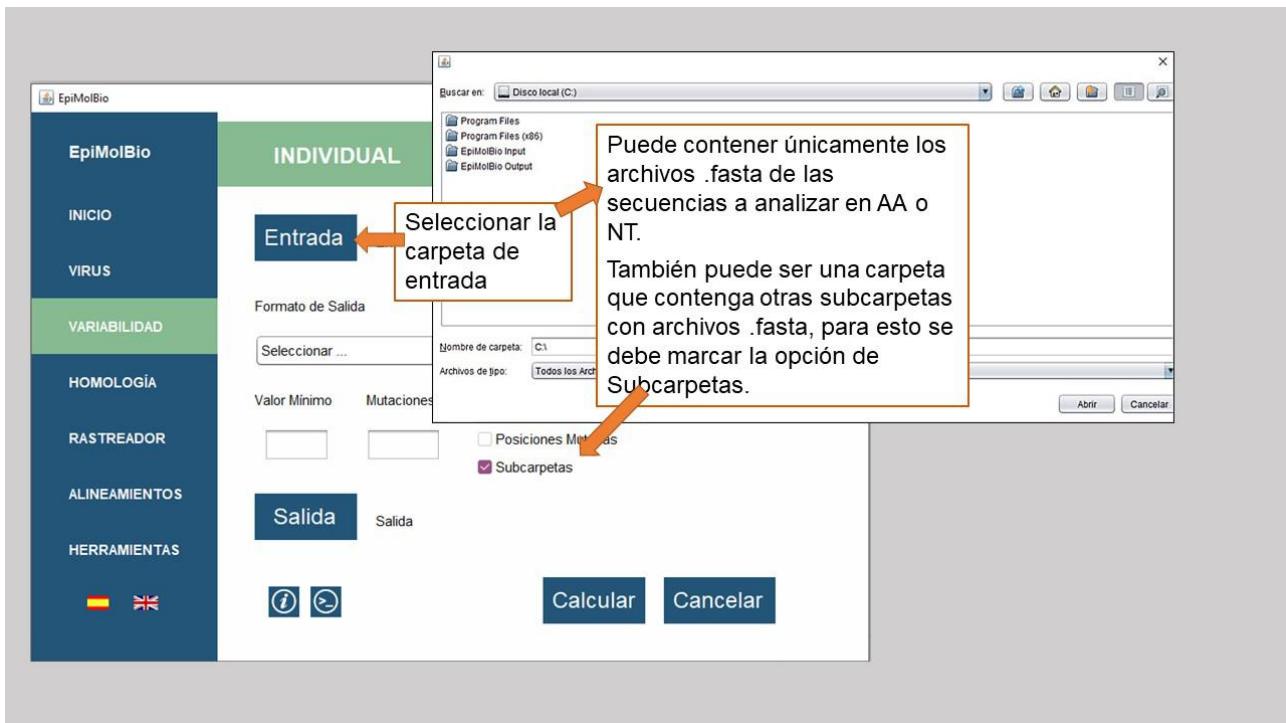
2)



3)



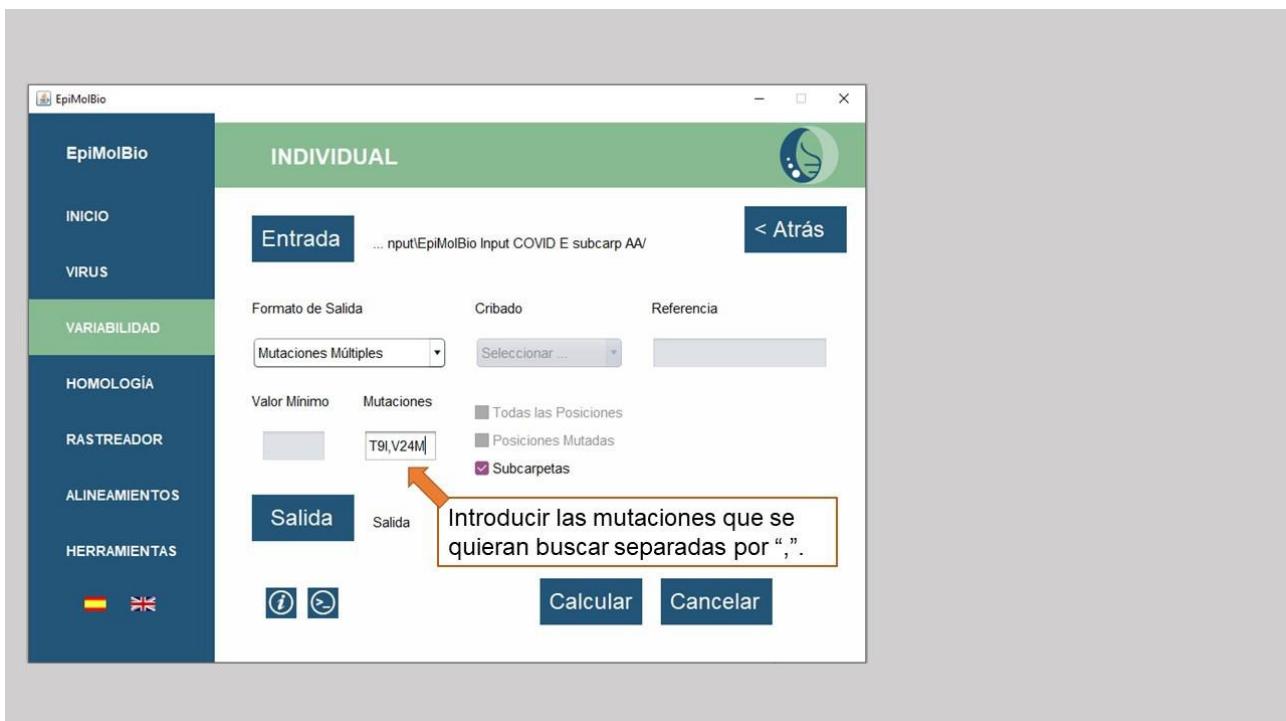
4)



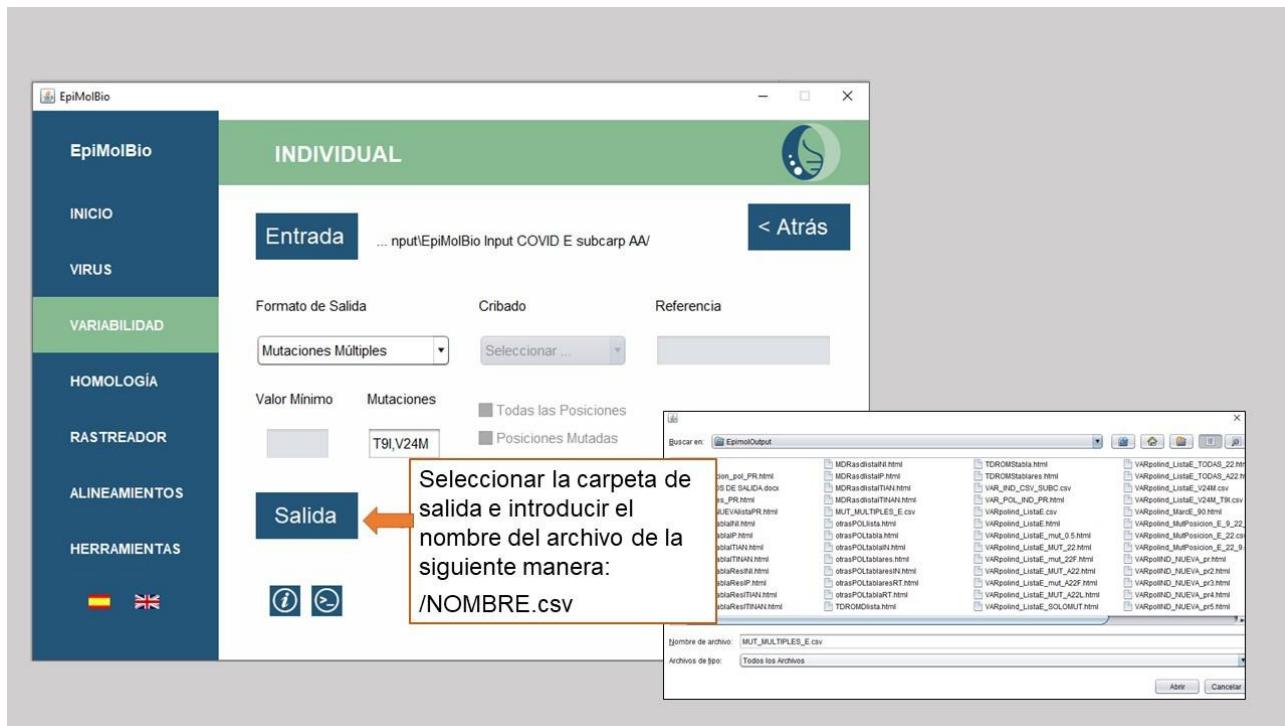
5)



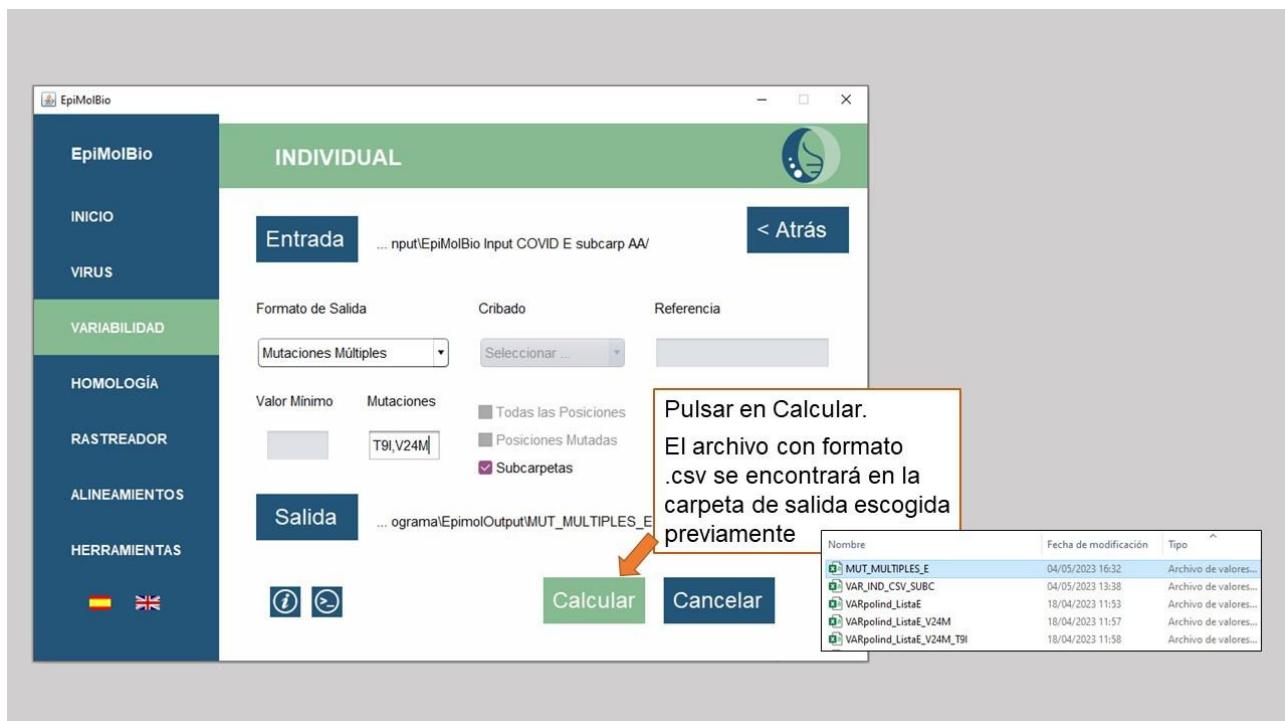
6)



7)



8)



5.-Mutaciones por Posición:

Permite **detectar y conocer la frecuencia de aparición de residuos** en una posición o varias posiciones combinadas, que deben introducirse previamente en el campo “Mutaciones”. De esta manera, se puede conocer qué residuos se encuentran en las posiciones de interés de nuestras secuencias y en qué combinaciones (ej.: buscar los residuos presentes en los tres sitios de unión de una molécula a la proteína de estudio).

El archivo .csv de salida es una tabla que contiene, en la primera columna, el nombre de los archivos de entrada; en la segunda, la combinación de residuos detectados en las posiciones introducidas en “Mutaciones”; en la tercera, el número de veces que aparece cada combinación; en la cuarta, la frecuencia de aparición de esa combinación de residuos y en la última columna, el número de secuencias válidas para esas posiciones. Sirve tanto para combinaciones de aminoácidos como de nucleótidos. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis. Si se quiere realizar el análisis en nucleótidos, se puede emplear la función “Buscar y reemplazar” en “Edición de Archivos” de “Herramientas” para cambiar las “N” por “?”, ya que la función no excluye las “N” del análisis.

Ejemplo de formato de salida Mutaciones por Posición introduciendo 3 posiciones:

| | A | B | C | D | E |
|---|----------------|---------------------|-------------------|------------|-------------------|
| 1 | Archivo | Residuos (12,15,17) | Número Mutaciones | Frecuencia | Número Secuencias |
| 2 | PR_01_AE.fasta | AIG | 289 | 1.134 | 25494 |
| 3 | PR_01_AE.fasta | AVG | 48 | 0.188 | 25494 |
| 4 | PR_01_AE.fasta | HIG | 1 | 0.004 | 25494 |
| 5 | PR_01_AE.fasta | IIG | 70 | 0.275 | 25494 |
| 6 | PR_01_AE.fasta | ILG | 1 | 0.004 | 25494 |
| 7 | PR_01_AE.fasta | IVG | 12 | 0.047 | 25494 |
| 8 | PR_01_AE.fasta | KIG | 13 | 0.051 | 25494 |

Para realizar este análisis, en **entrada** se debe seleccionar una carpeta donde se tengan exclusivamente los archivos en formato. Fasta en aminoácidos o nucleótidos, o una carpeta que contenga otras subcarpetas con archivos .fasta. Para ésto se debe marcar la opción de Subcarpetas.

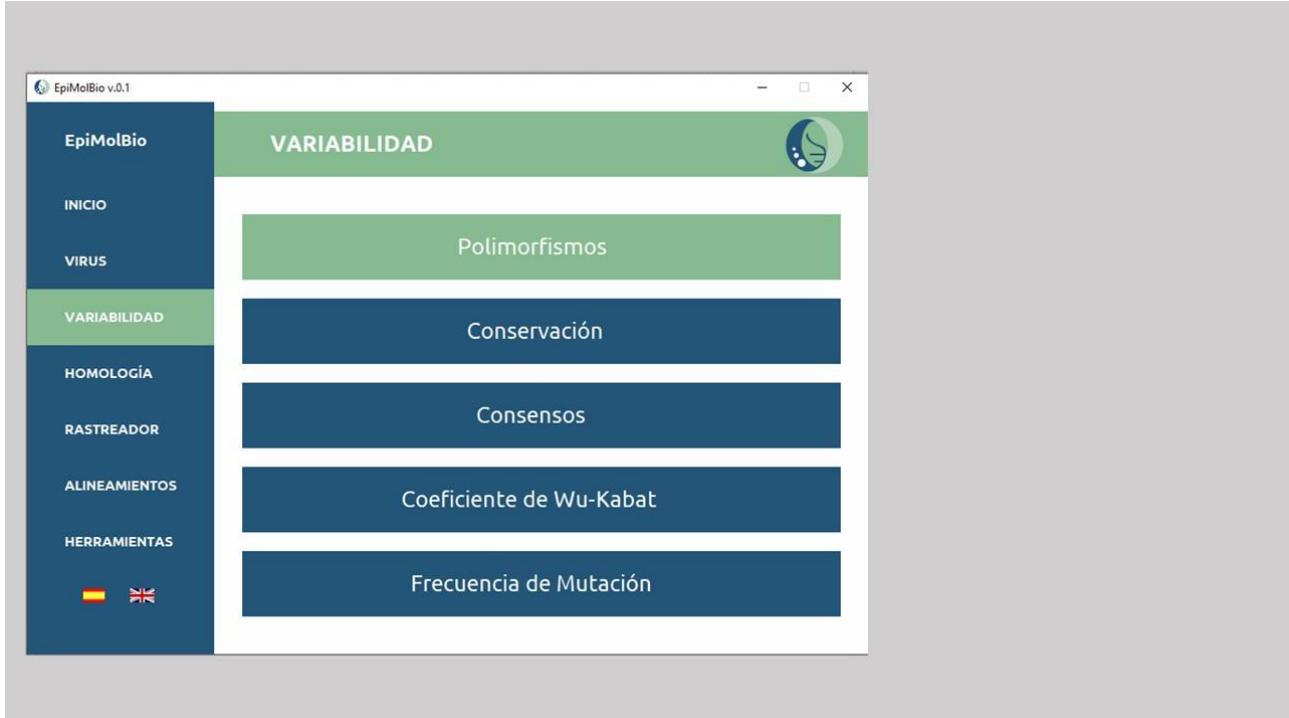
Habrá que seleccionar “**Mutaciones por posición**” en el desplegable “**Formato de salida**”.

En el campo “**Mutaciones**”, introducir las posiciones que se desean analizar tecleando el número de cada posición separadas por una coma “,” sin espacios (ej.: 9,22,30).

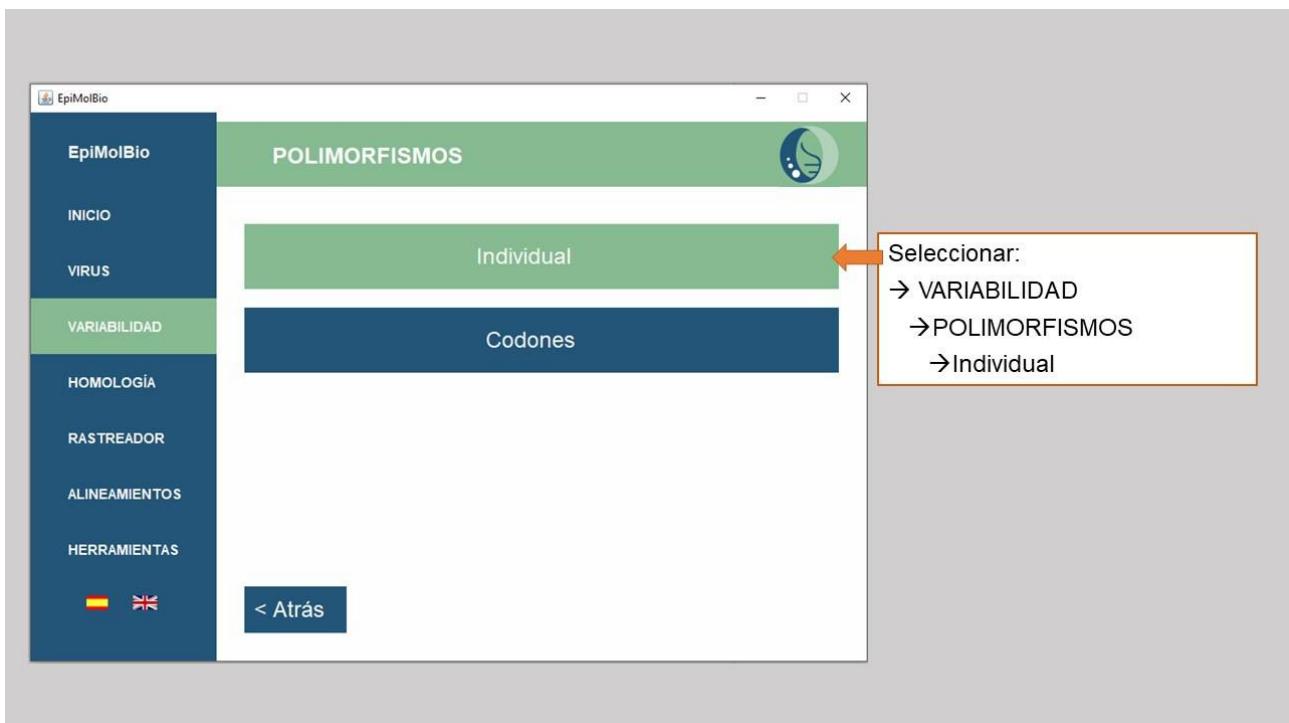
El resultado se muestra en un archivo .csv. En **salida** se debe seleccionar la carpeta donde se quiera guardar el resultado y nombrar el archivo con la extensión .csv.

Paso a paso:

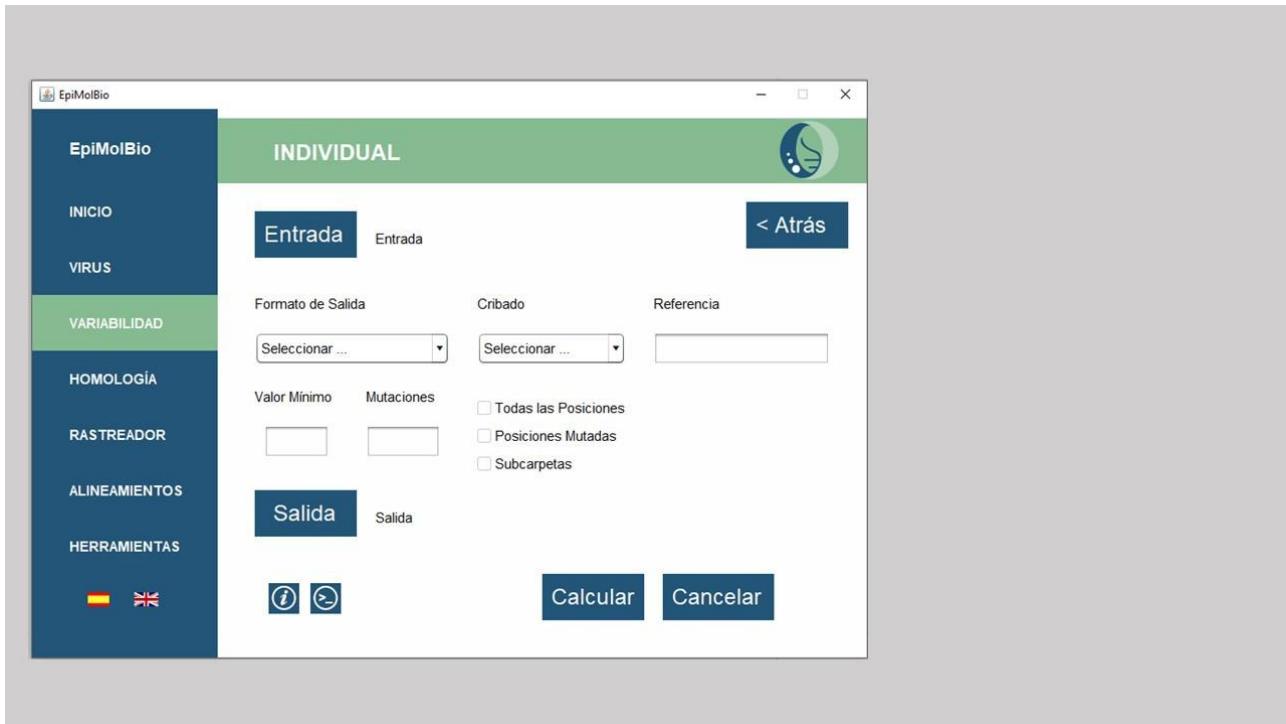
1)



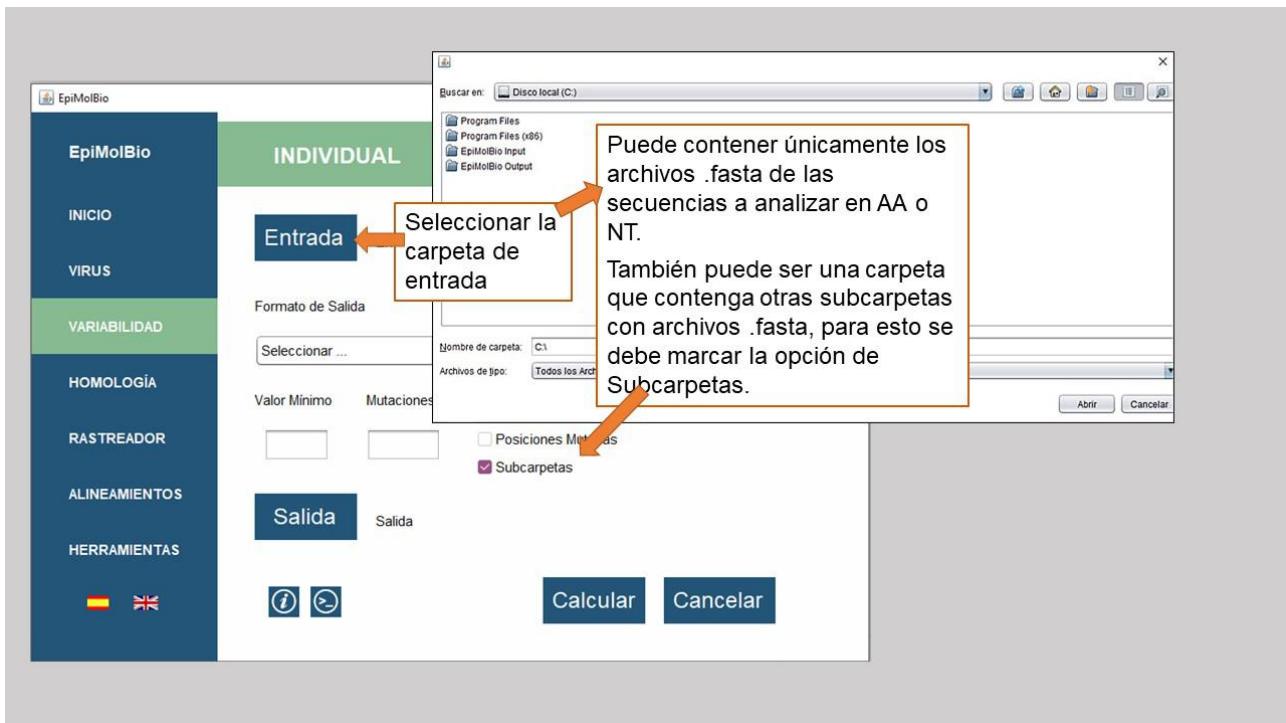
2)



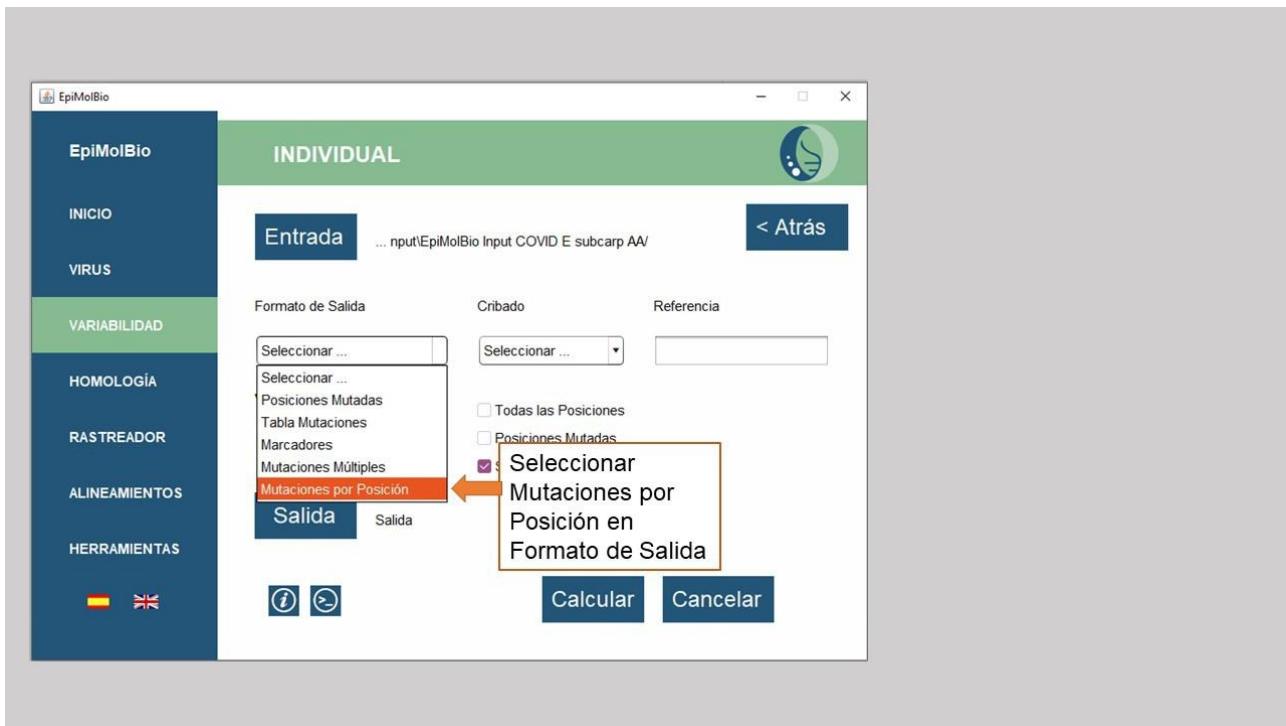
3)



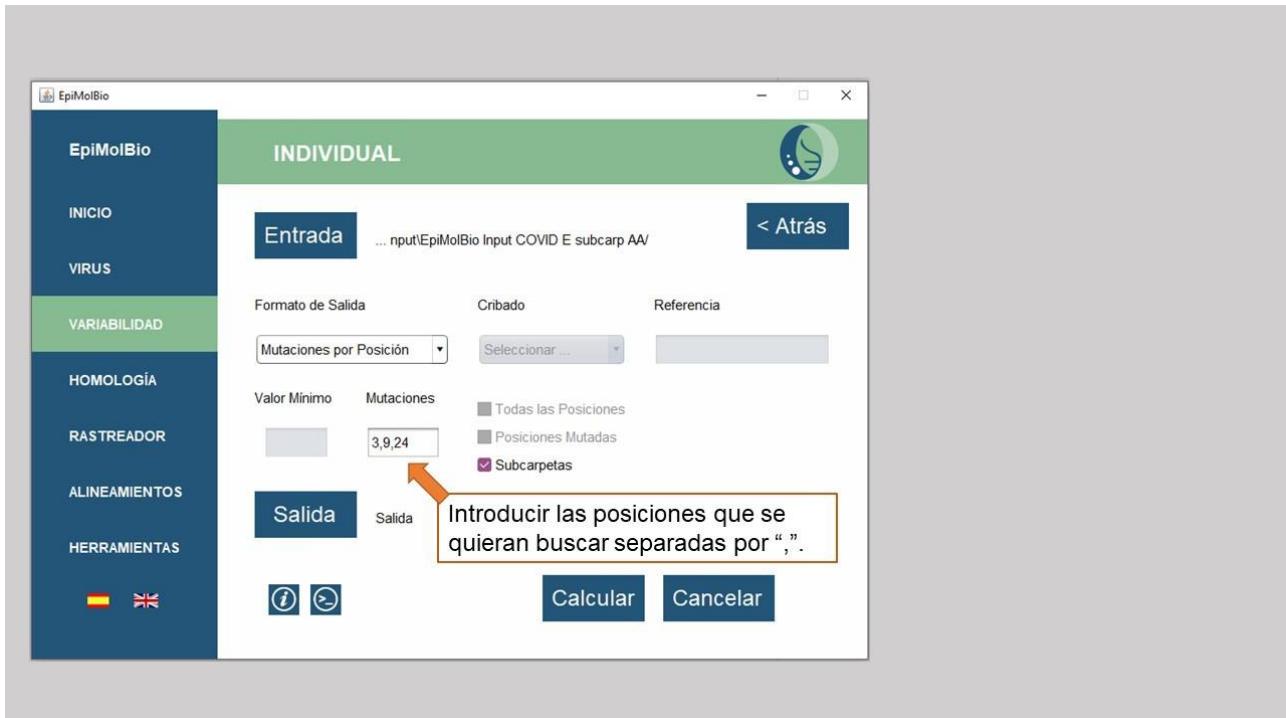
4)



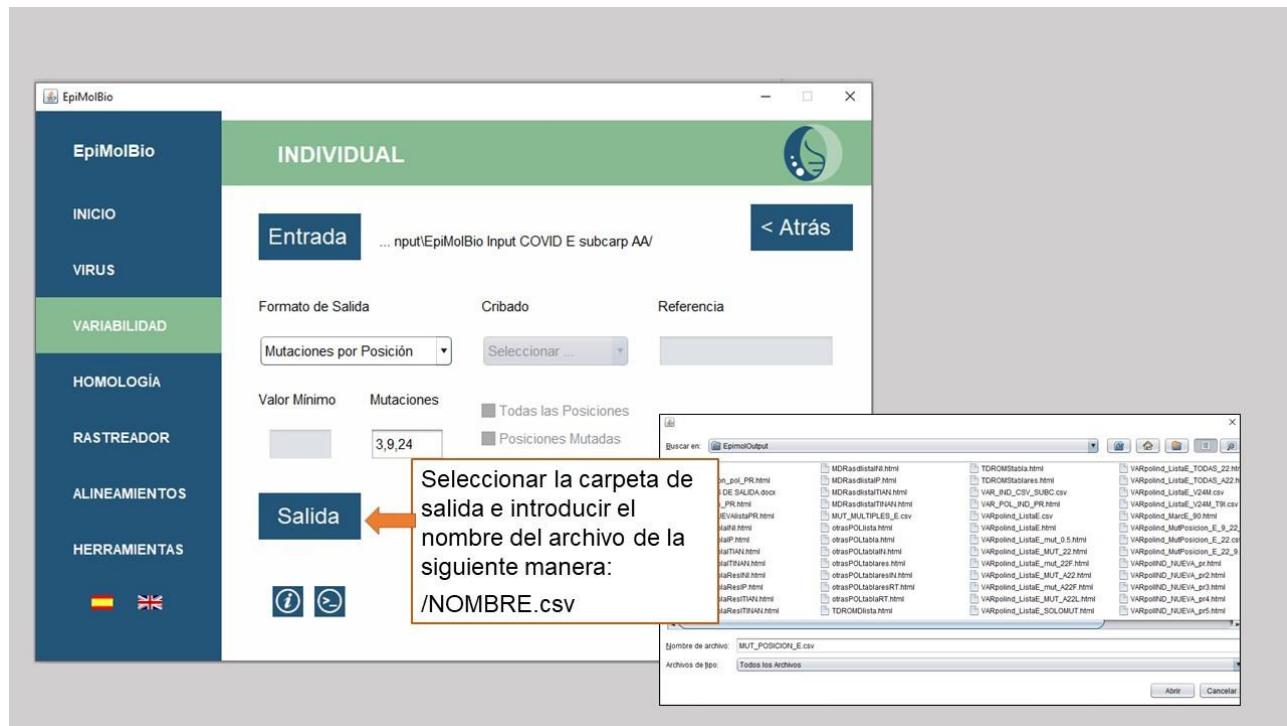
5)



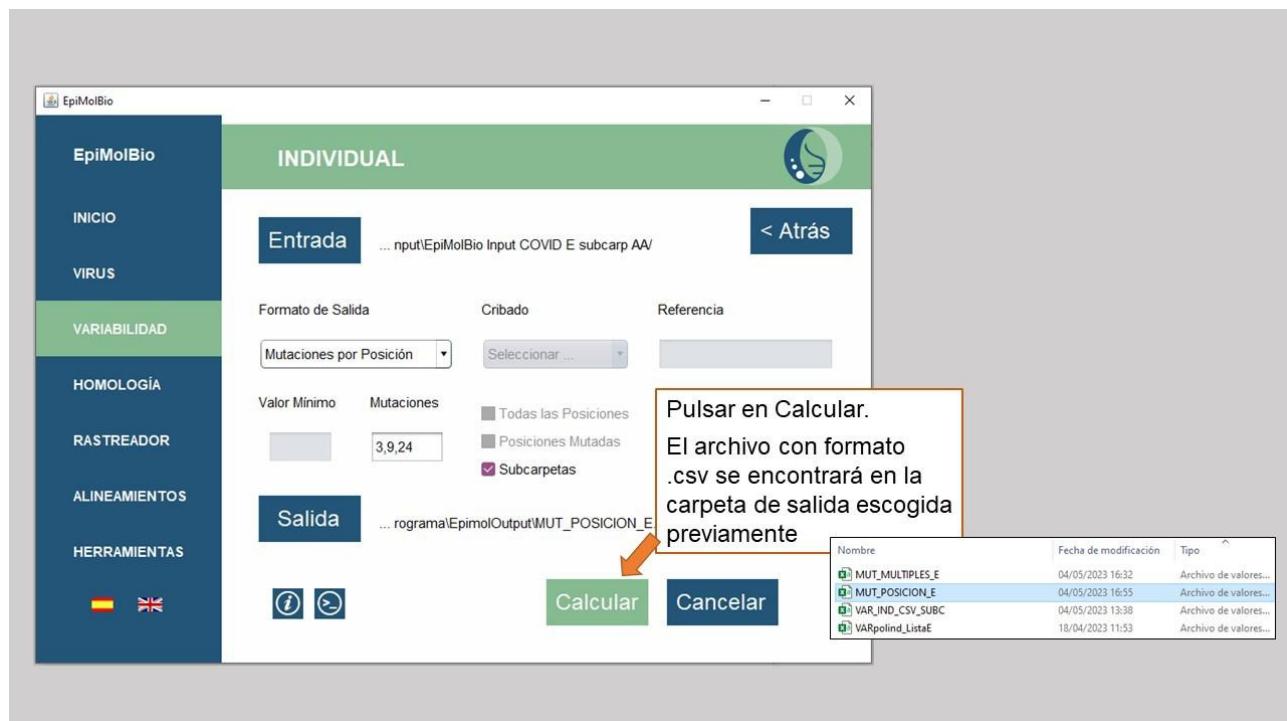
6)



7)



8)



II.1.B) CODONES

Esta función permite detectar todos los codones que son diferentes a los de la secuencia de referencia y su frecuencia de aparición a partir de secuencias de nucleótidos en formato .fasta utilizando una secuencia de referencia introducida por el usuario. Los gaps (-) y las "?" son excluidos del análisis.

El formato del archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar en nucleótidos.

En el campo “**Cribado**” se puede seleccionar la frecuencia de aparición mínima para la detección de codones, escogiendo entre 100% para que detecte todos los codones distintos a los de la secuencia de referencia introducida, o >75% para que detecte sólo aquellos que se presenten con una frecuencia superior al 75%.

En “**Referencia**” introducir la secuencia de referencia en letras sin saltos de línea, teniendo en cuenta que la secuencia debe introducirse en nucleótidos.

El formato de **salida** será un archivo con extensión .html. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

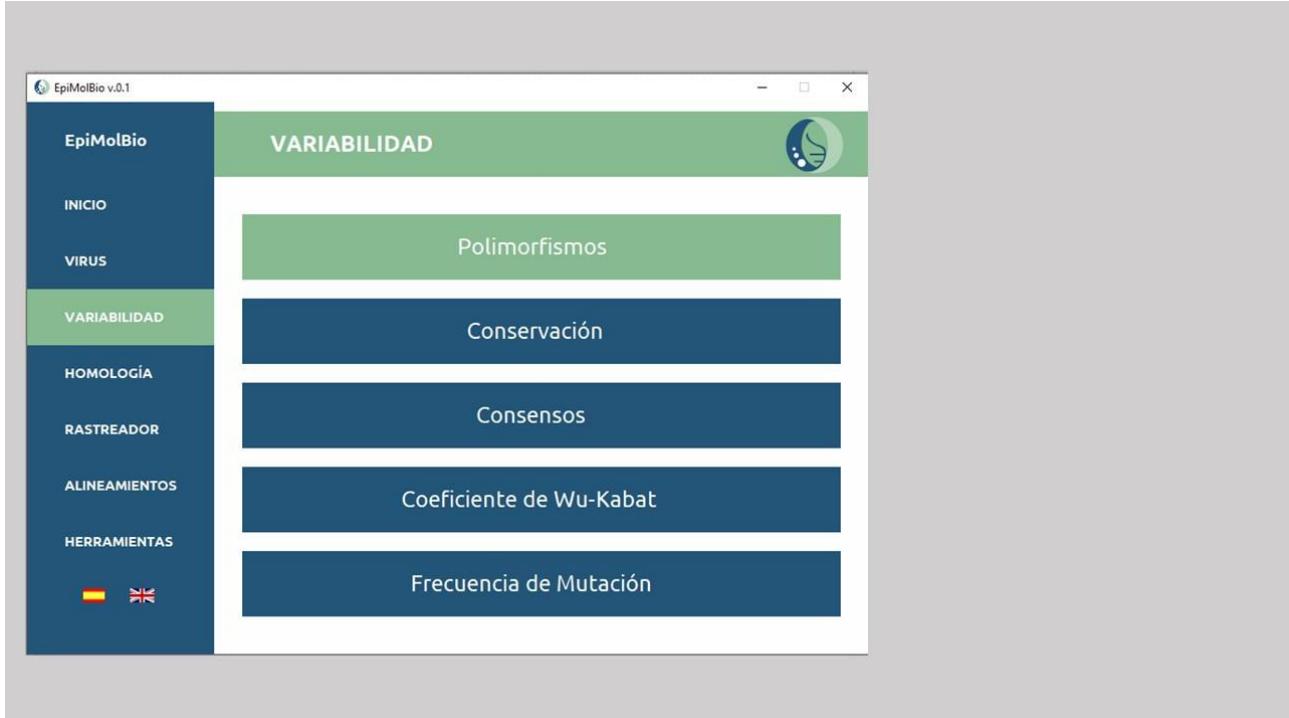
En el archivo de salida aparece, en la parte superior, el título del análisis seguido del nombre del archivo de entrada. Debajo, en la columna “Posición”, aparece el codón analizado correspondiente a la secuencia de referencia con el aminoácido que codifica y los nucleótidos entre paréntesis. En la columna “Residuos” aparecen todas las variaciones detectadas con el siguiente formato: aminoácido codificado [codón detectado](frecuencia de aparición coloreada según el código de colores descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul)]. En la columna “Posiciones Totales”, aparece el número total de secuencias válidas para ese codón. Si uno de los codones no presenta variaciones, no aparecerá en el archivo de salida.

Ejemplo del archivo de salida del análisis de Polimorfismos Codones:

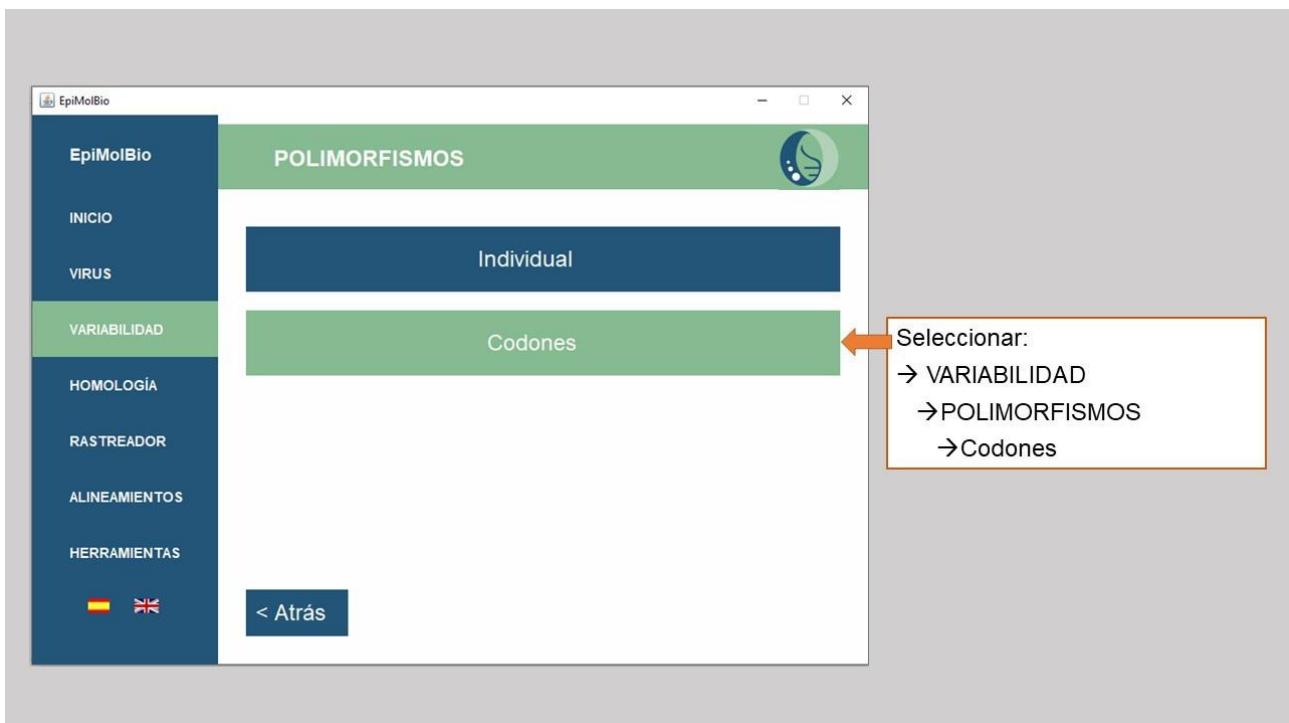
| Variabilidad Polimorfismos Codones > 75% | | |
|--|--------------------------|--------------------|
| PR_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| 2 Q(CAG) | Q[CAA(96.554%)] | 26844 |
| 3 V(GTC) | I[ATC(99.765%)] | 26847 |
| 10 L(CTC) | L[CTT(78.294%)] | 26840 |
| 14 K(AAG) | K[AAA(91.358%)] | 26845 |
| 17 G(GGG) | G[GGA(92.468%)] | 26845 |
| 18 Q(CAA) | Q[CAG(89.655%)] | 26845 |
| 35 E(GAA) | D[GAT(80.933%)] | 26847 |
| 36 M(ATG) | I[ATA(98.678%)] | 26846 |
| 37 S(AGT) | N[AAT(90.951%)] | 26844 |

Paso a paso:

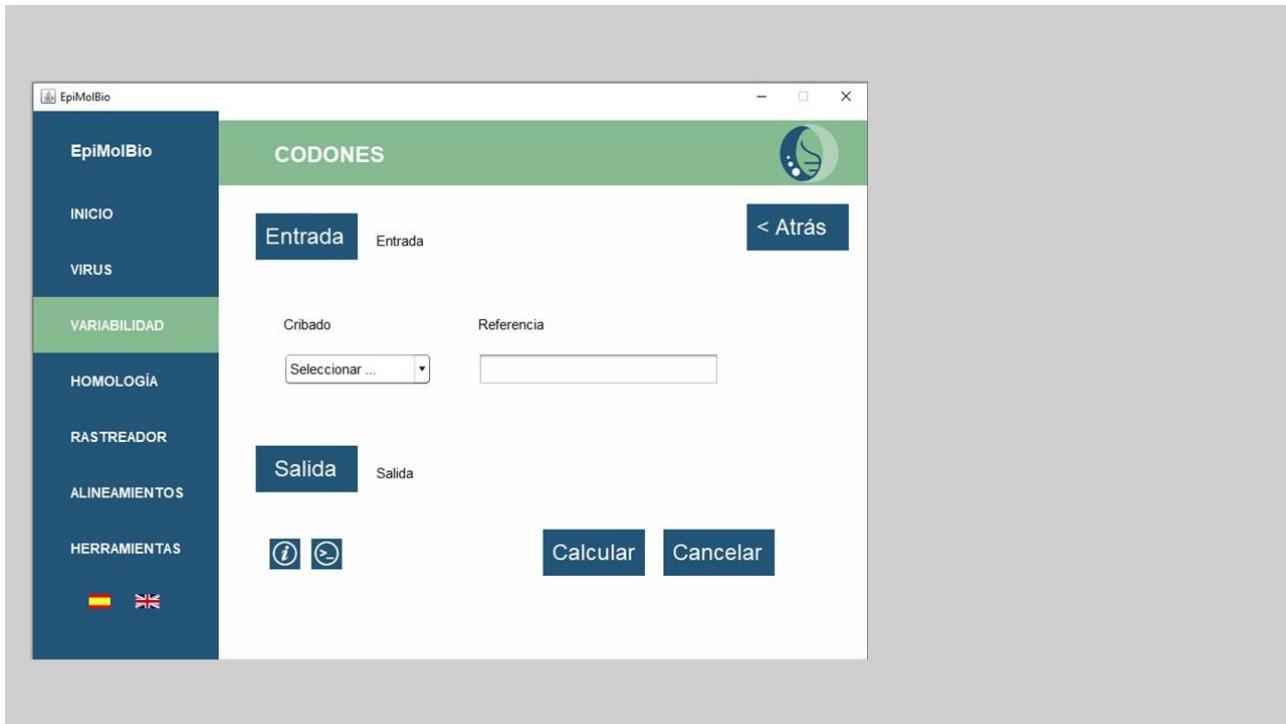
1)



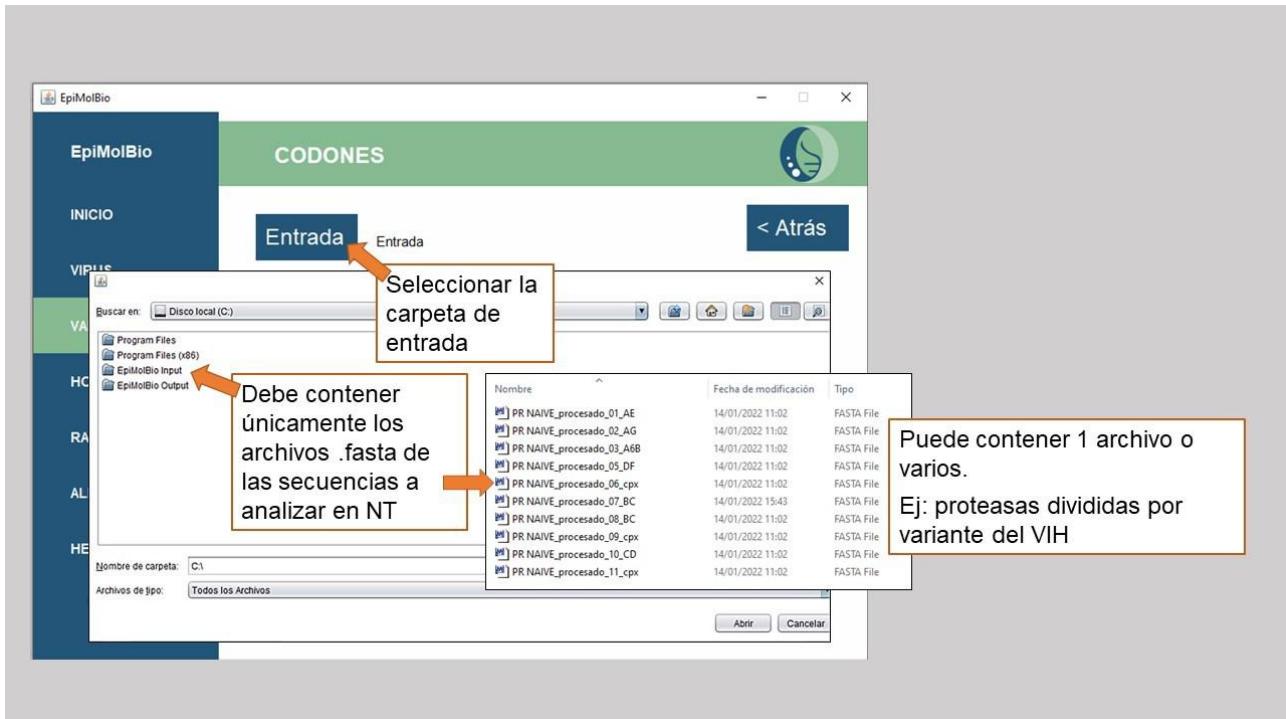
2)



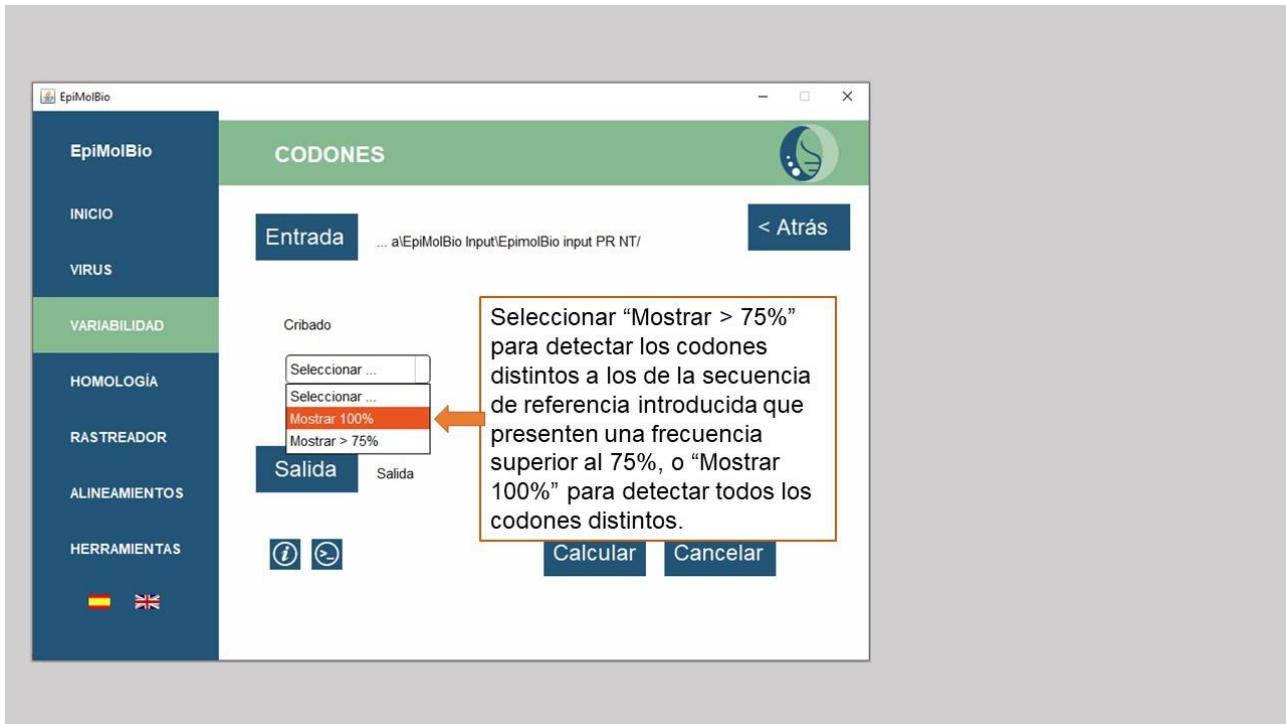
3)



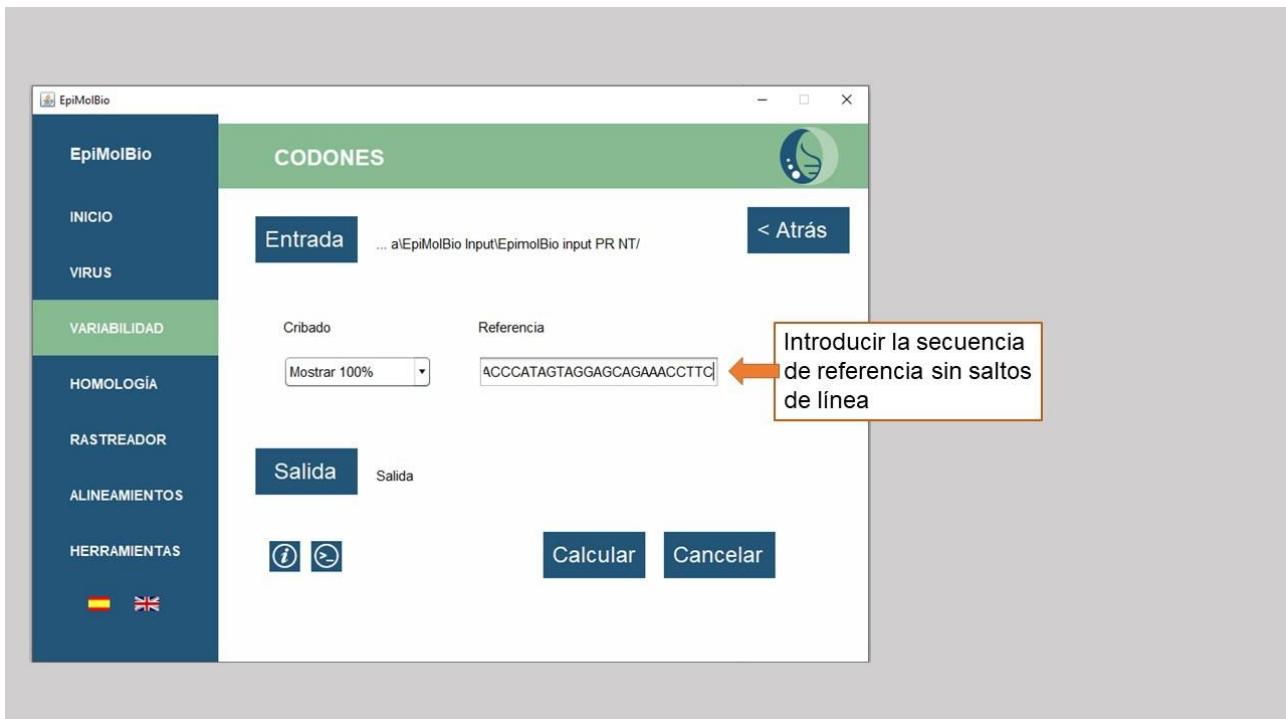
4)



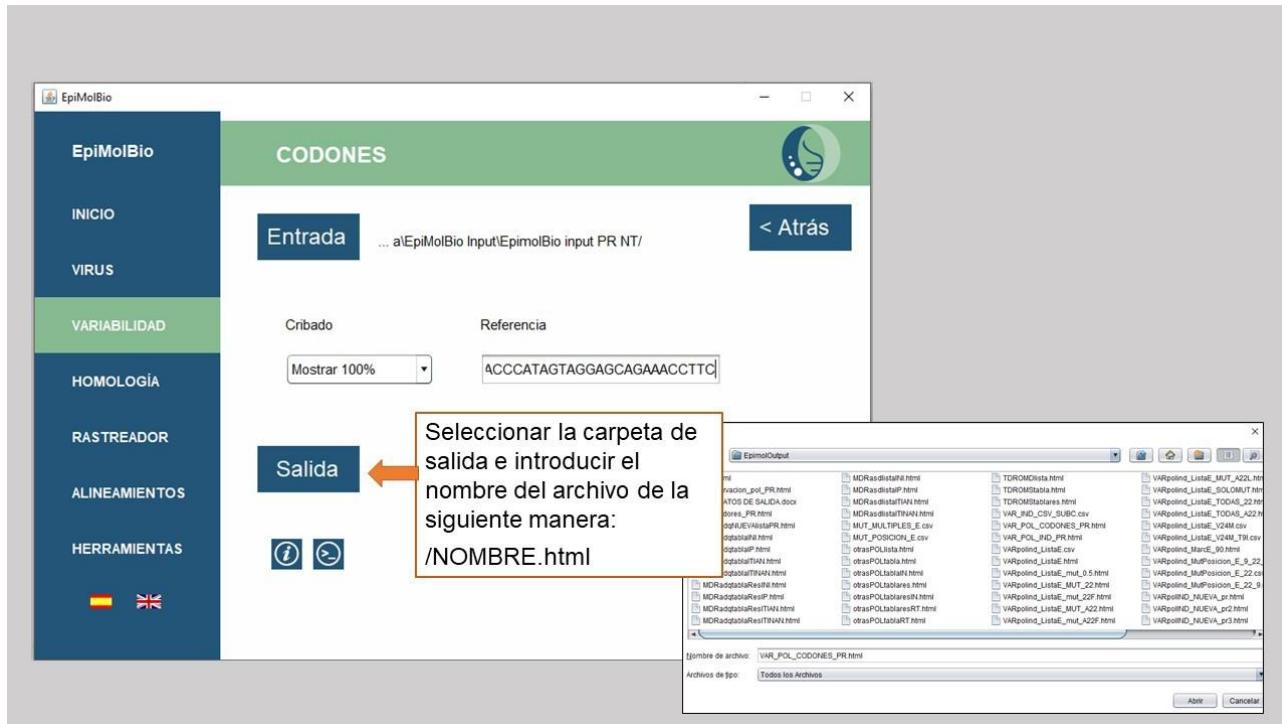
5)



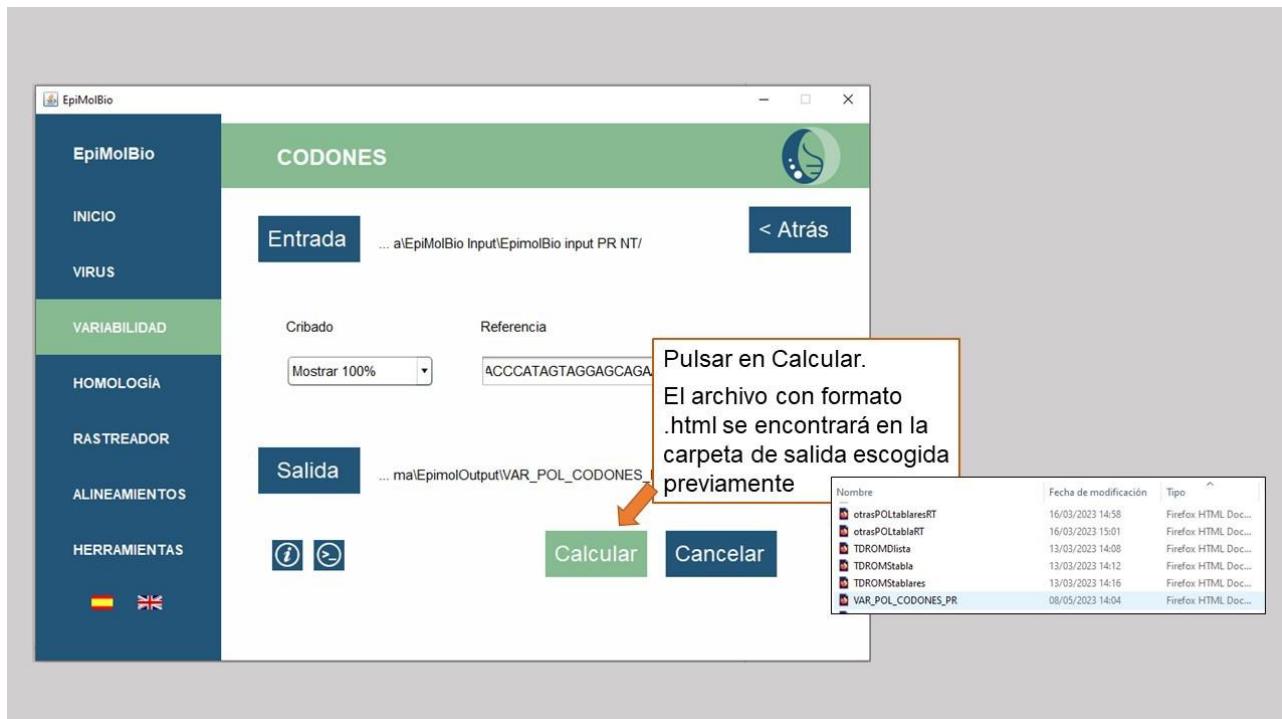
6)



7)



8)



II.2. CONSERVACIÓN

Esta función permite conocer el grado de conservación de secuencias de interés informando del residuo o codón más prevalente y su porcentaje. También permite aplicar cribados para acotar las frecuencias y generando secuencias consenso a partir de los archivos de entrada.

II.2.A) INDIVIDUAL

Permite obtener el aminoácido o nucleótido más conservado para cada posición de la secuencia analizada. Tanto los gaps (-) como las interrogaciones (?) son excluidas del análisis. En algunos formatos de salida, según el cribado escogido, puede ocurrir que dos residuos aparezcan como los más conservados porque presenten la misma frecuencia de aparición, en estos casos, se mostrarán ambos residuos en el resultado.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar alineadas. Las secuencias pueden ser tanto de nucleótidos como de aminoácidos. Para realizar el análisis en secuencias de nucleótidos será necesario emplear la herramienta Buscar y Reemplazar de Edición de Archivos y sustituir las “N” por “?” para que estas se excluyan del análisis.

En **Formato de Salida** se puede escoger entre dos formatos de resultados diferentes: Lista y Tabla. Todos ellos tienen la extensión .html. El formato **Lista** utiliza por defecto el **cribado >75%**, por lo que sólo se mostrarán los residuos conservados que presenten una frecuencia mayor del 75%. El formato **Tabla** utiliza por defecto el **cribado 100%**, por lo que se mostrarán todos los residuos totalmente conservados.

En el campo “**Referencia**” se debe introducir una secuencia de referencia sin saltos de línea, en NT o AA según el archivo de entrada.

El archivo de **salida** será un archivo con extensión .html. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

1.-Lista:

En el formato de salida Lista aparece, en la parte superior, el título del análisis seguido del nombre del archivo de entrada. Debajo, se muestra la secuencia consenso para cada archivo con los residuos coloreados según su porcentaje siguiendo el código de colores descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul. En la columna “Posición” aparecen las posiciones que contengan un residuo conservado con una frecuencia superior al 75% con su aminoácido de referencia. En la columna “Residuos” aparece el aminoácido más frecuente para cada posición seguido de su porcentaje coloreado según el código de colores. En la columna “Posiciones Totales” aparece el número total de secuencias válidas para esa posición.

Ejemplo de formato de salida Lista para el análisis de Conservación Individual:

| Variabilidad Lista Consevación Individual | | |
|---|---|--------------------|
| PR_01_AE.fasta | | |
| CONSENSO | PQITLWQRPLVTVKIGQLKEALLDTGADDVLEDINLPGKWKPKMIGGIGGFIKVHQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | |
| Posición | Residuos | Posiciones Totales |
| | P(99.896%) | 26838 |
| | Q(99.782%) | 26649 |
| | I(99.858%) | 26831 |
| | T(99.858%) | 26816 |
| | L(99.888%) | 26780 |
| | W(99.929%) | 26836 |
| Q7 | Q(99.751%) | 26536 |

2.-Tabla:

En el formato de salida Tabla aparece, en la parte superior, el título del análisis. Debajo, en las primeras dos filas, se muestra la secuencia de referencia y las posiciones de cada uno de sus residuos. El análisis aparece en las siguientes filas mostrando el archivo de entrada y tres filas correspondientes al residuo más frecuente (nucleótido o aminoácido) con la celda coloreada siguiendo el código de colores previamente descrito, la frecuencia de conservación por posición también con la celda coloreada y el número de secuencias válidas por cada posición.

Ejemplo de formato de salida Tabla para el análisis de Conservación Individual:

| Variabilidad Tabla Conservación Individual | | | | | | | | | | | | | | | | | | |
|--|----------------------|---------|---------|---------|--------|---------|---------|--------|--------|--------|--------|---------|--------|--------|--------|--------|--------|--|
| Archivo | Referencia | P | Q | V | T | L | W | Q | R | P | L | V | T | I | K | I | G | |
| | Posición | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| PR_01_AE.fasta | Residuo | P | Q | I | T | L | W | Q | R | P | L | V | T | V | K | I | G | |
| | Conservación | 99.896 | 99.782 | 99.858 | 99.858 | 99.888 | 99.929 | 99.751 | 99.922 | 99.955 | 87.446 | 99.512 | 95.028 | 56.099 | 95.188 | 88.392 | 72.550 | |
| | Número de Secuencias | 26838 | 26649 | 26831 | 26816 | 26780 | 26836 | 26536 | 26792 | 26613 | 25952 | 26416 | 26469 | 26150 | 26270 | 26206 | 25636 | |
| PR_02_AG.fasta | Residuo | P | Q | I | T | L | W | Q | R | P | L | V | T | V | R | I | G | |
| | Conservación | 99.948 | 99.819 | 99.529 | 99.728 | 99.958 | 99.875 | 99.838 | 99.895 | 99.947 | 83.181 | 96.532 | 89.274 | 91.362 | 59.803 | 85.337 | 70.800 | |
| | Número de Secuencias | 9571 | 9416 | 9557 | 9561 | 9560 | 9575 | 9248 | 9557 | 9351 | 9186 | 9313 | 9295 | 9308 | 8916 | 9289 | 9233 | |
| PR_03_A6B.fasta | Residuo | P | Q | I | T | L | W | Q | R | P | L | V | T | V | K | I | G | |
| | Conservación | 100.000 | 100.000 | 100.000 | 99.677 | 100.000 | 100.000 | 99.020 | 99.676 | 99.672 | 85.284 | 100.000 | 91.803 | 58.020 | 57.241 | 85.714 | 74.074 | |
| | Número de Secuencias | 310 | 307 | 310 | 310 | 310 | 310 | 306 | 309 | 305 | 299 | 301 | 305 | 293 | 290 | 301 | 297 | |

Paso a paso:

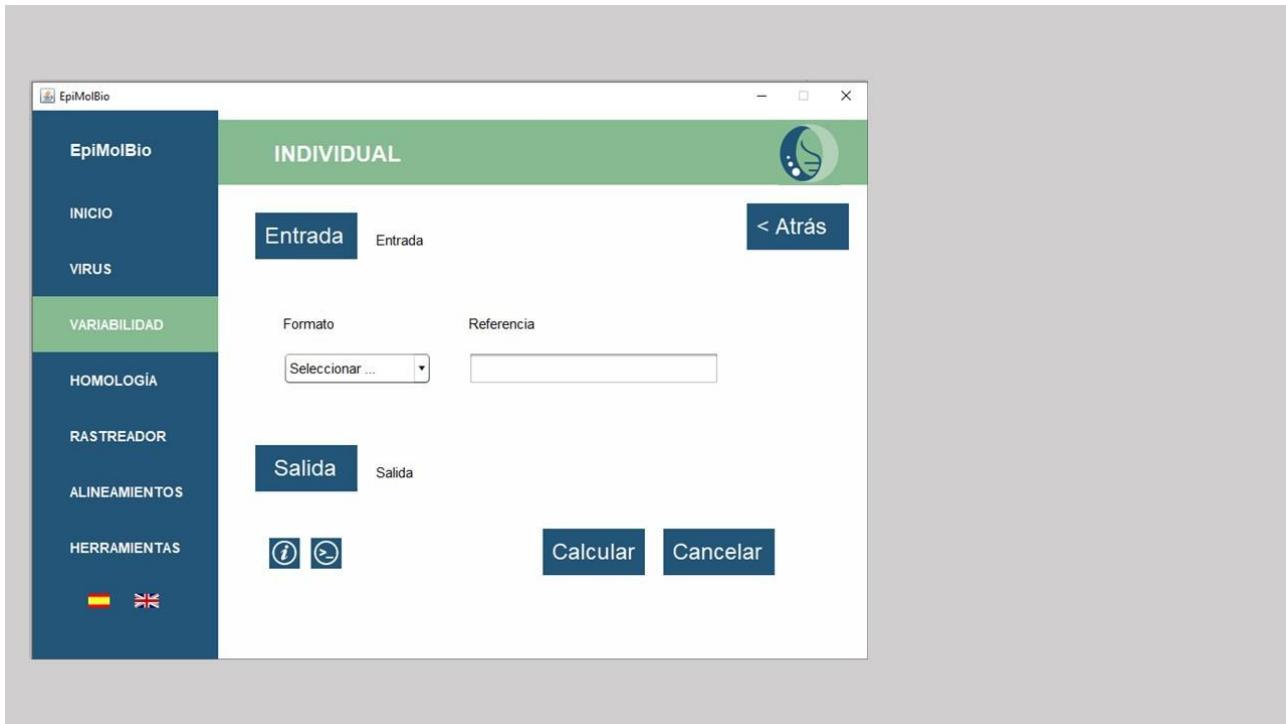
1)



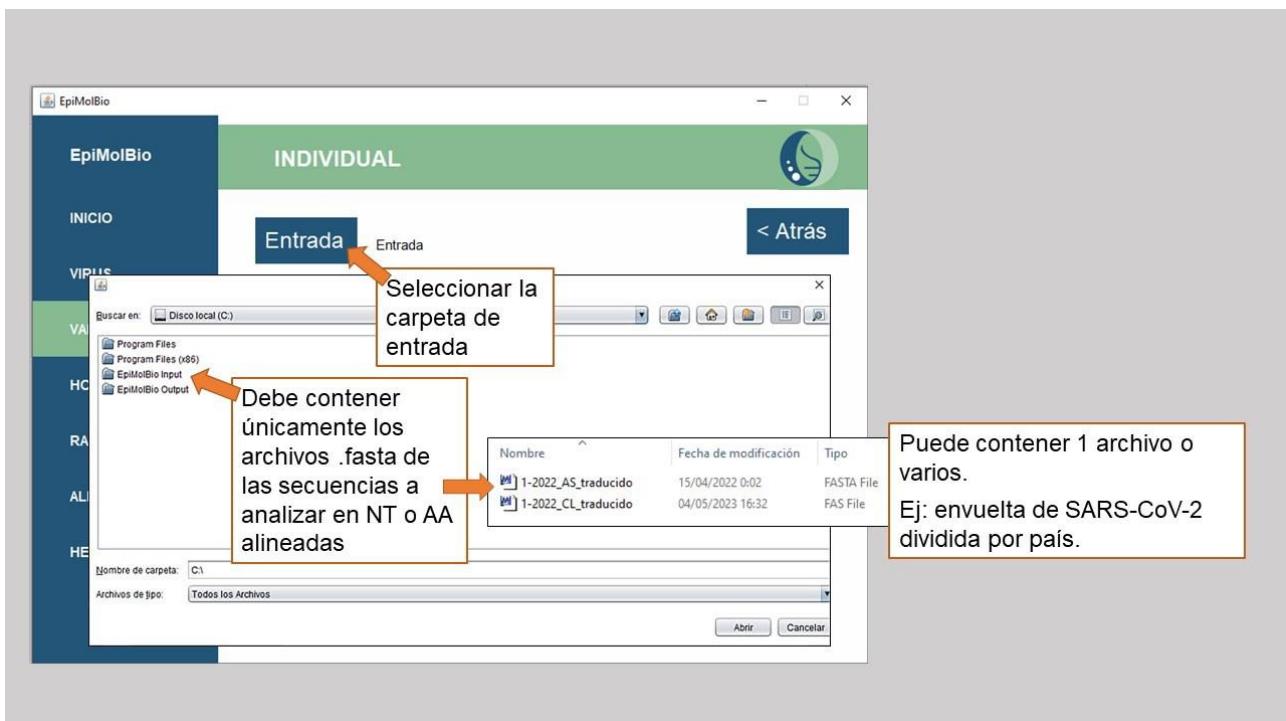
2)



3)



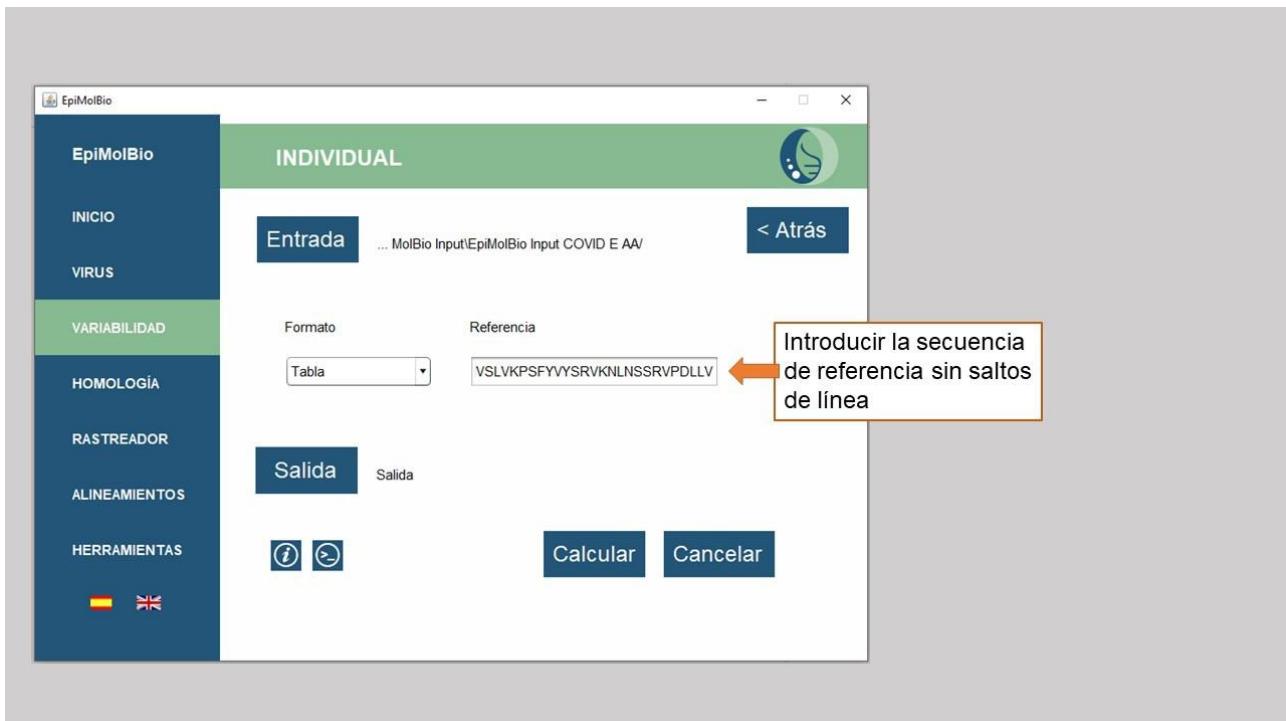
4)



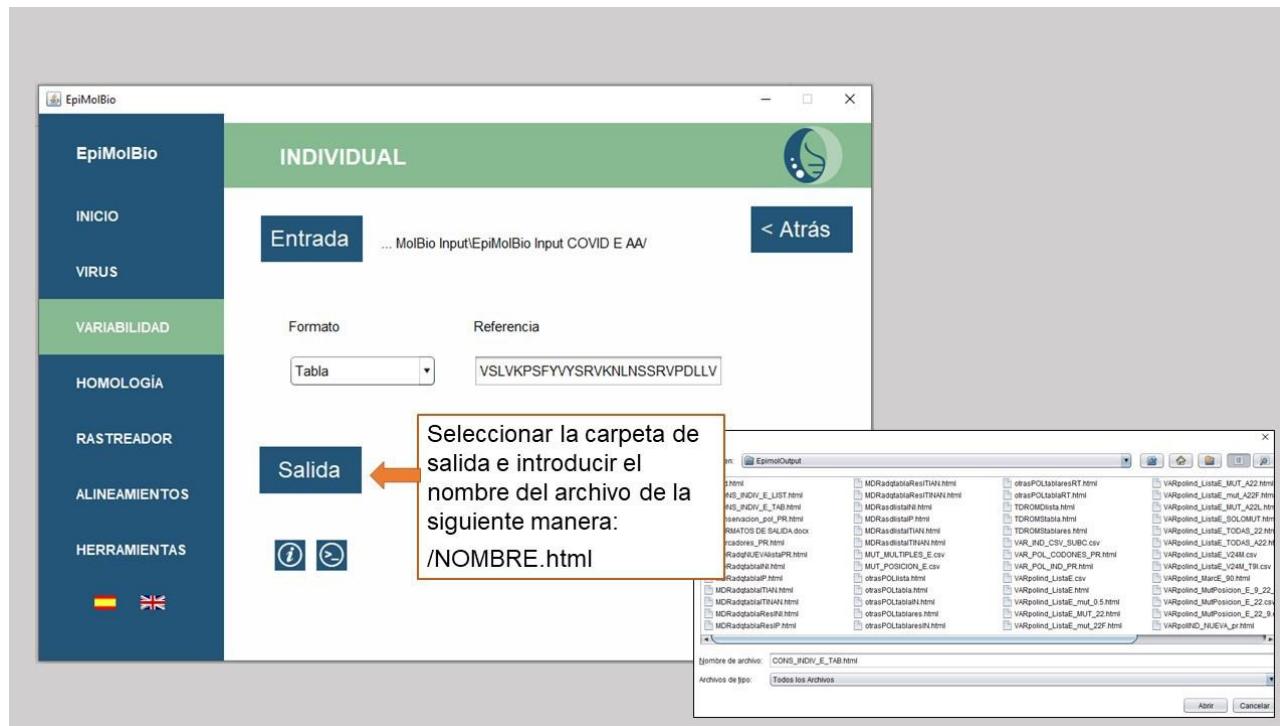
5)



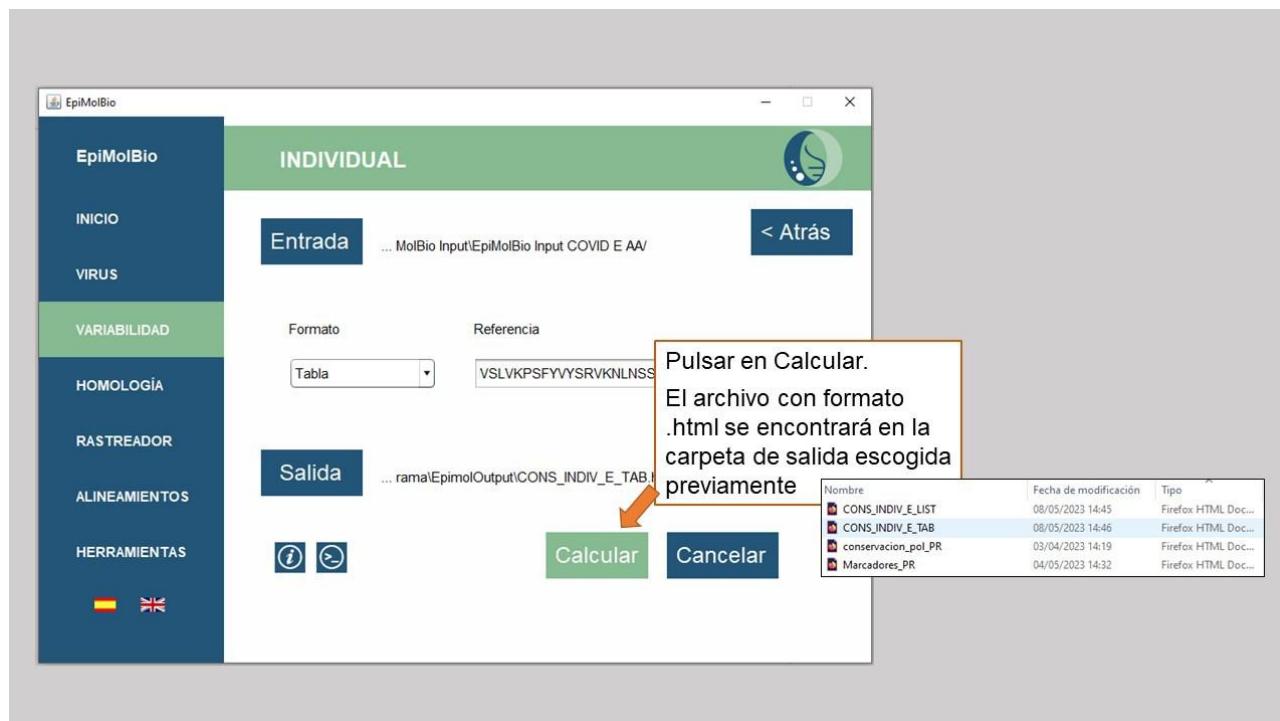
6)



7)



8)



II.2.B) CODONES

Esta función **permite obtener el codón más conservado para cada triplete** de una secuencia de nucleótidos analizada. Tanto los gaps (-) como las “N” son excluidas del análisis.

El formato del archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar alineadas. Las secuencias deben ser de NT.

En el campo “**Cribado**” escoger el porcentaje de cribado: **100%** para que se muestren todos los codones encontrados, o **>75%** para que se muestren sólo aquellos con una frecuencia de aparición superior al 75%.

En el campo “**Referencia**” se debe introducir una secuencia de referencia sin saltos de línea y en nucleótidos.

El formato de **salida** será un archivo con extensión .html. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

En el archivo de salida aparece, en la parte superior, el título del análisis seguido del nombre del archivo de entrada. En la columna “Posición”, la posición del aminoácido codificado por cada codón analizado con el codón entre paréntesis. En la columna “Residuos”, todos los residuos encontrados [el codón y su porcentaje] si el cribado escogido es el 100%. Si se escoge el cribado >75%, sólo se mostrará el codón más conservado que aparezca con una frecuencia > 75% en las secuencias analizadas. Los porcentajes estarán coloreados según el **código de colores** descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul. En la columna “Posiciones Totales” aparece el número total de secuencias válidas para esa posición.

Ejemplo de formato de salida con cribado 100%:

| Variabilidad Conservación Codones 100% | | |
|--|---|--------------------|
| PR_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| 1 P(CCT) | P[CCT(99.810%)] S[CTT(0.011%)] P[CCA(0.011%)] ?[CCY(0.022%)] P[CCC(0.026%)] S[TCT(0.067%)] A[GCT(0.004%)] L[CTT(0.004%)] ?[SCT(0.007%)] ?[CCW(0.004%)] P[CCG(0.007%)] T[ACT(0.007%)] H[CAT(0.004%)] ?[YCT(0.004%)] ?[CMT(0.004%)] V[GTC(0.004%)] L[CTC(0.004%)] | 26849 |
| 2 Q(CAG) | Q[CAA(96.554%)] Q[CAG(2.503%)] E[GAA(0.071%)] ?[CAR(0.596%)] S[ATCA(0.019%)] H[CAT(0.034%)] D[GAC(0.004%)] K[AAA(0.022%)] L[CTG(0.004%)] H[CAC(0.022%)] ?[CAM(0.034%)] ?[CWW(0.004%)] ?[YAA(0.004%)] ?[CAW(0.026%)] ?[CMM(0.015%)] ?[SAA(0.007%)] P[CCT(0.004%)] P[CCC(0.004%)] L[CTC(0.011%)] ?[CWM(0.004%)] R[CGA(0.004%)] R[CGC(0.004%)] T[ACA(0.004%)] ?[CRA(0.019%)] ?[CMA(0.004%)] *(TAG(0.007%)] ?[CAV(0.004%)] ?[MAR(0.004%)] ?[CWA(0.007%)] R[AGA(0.004%)] | 26844 |

Ejemplo de formato de salida con cribado 75%:

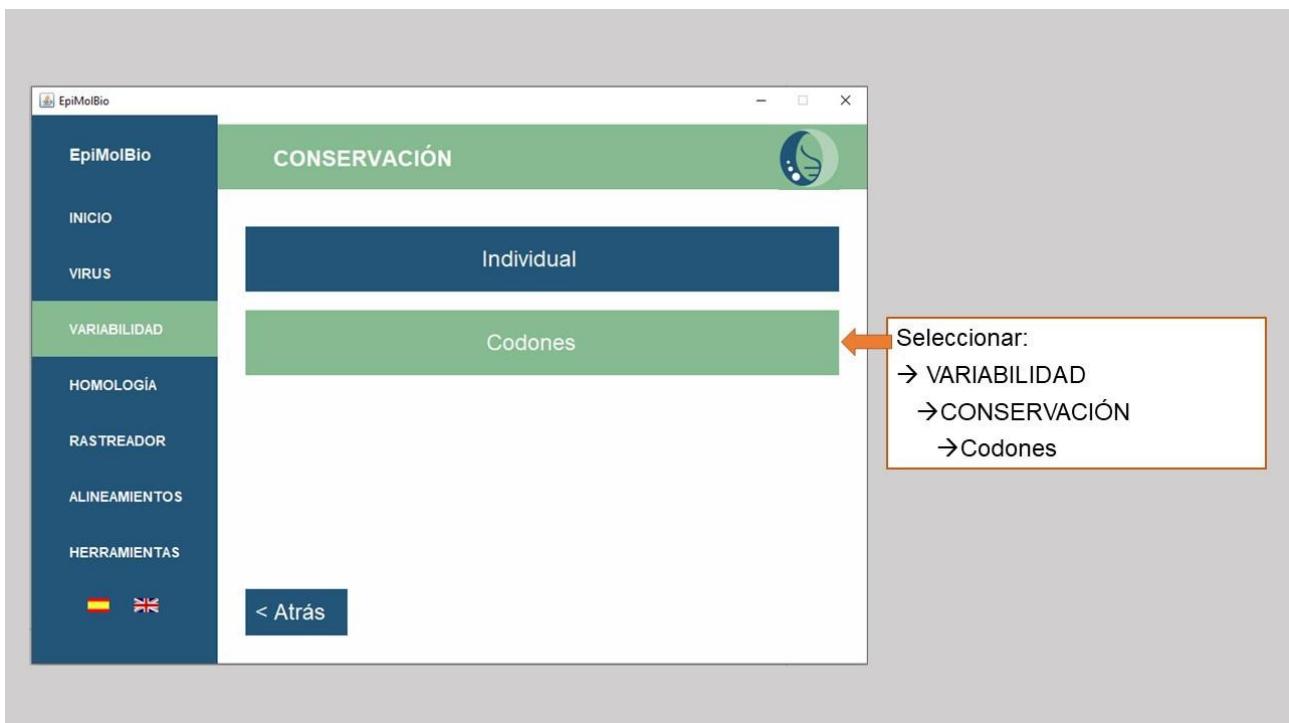
| Variabilidad Conservación Codones > 75% | | |
|---|-----------------|--------------------|
| PR_01_AE.fasta | | |
| Posición | Residuos | Posiciones Totales |
| 1 P(CCT) | P[CCT(99.810%)] | 26849 |
| 2 Q(CAG) | Q[CAA(96.554%)] | 26844 |
| 3 V(GTC) | I[ATC(99.765%)] | 26847 |

Paso a paso:

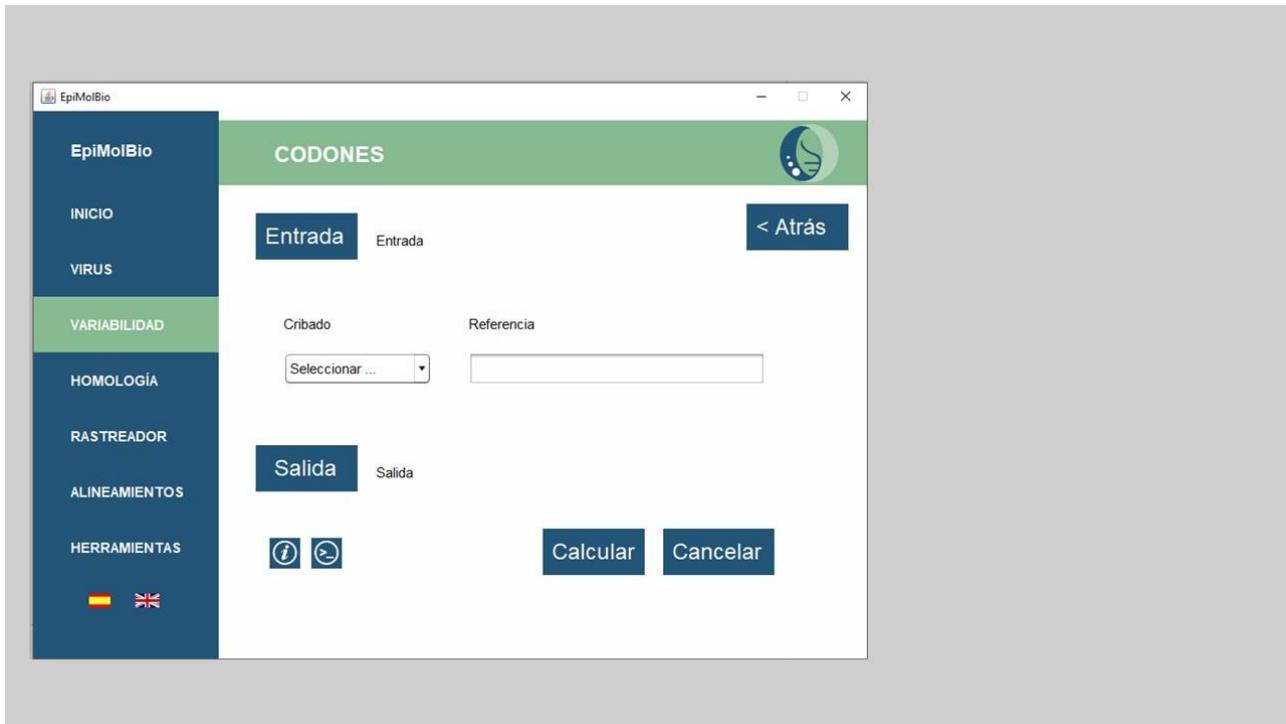
1)



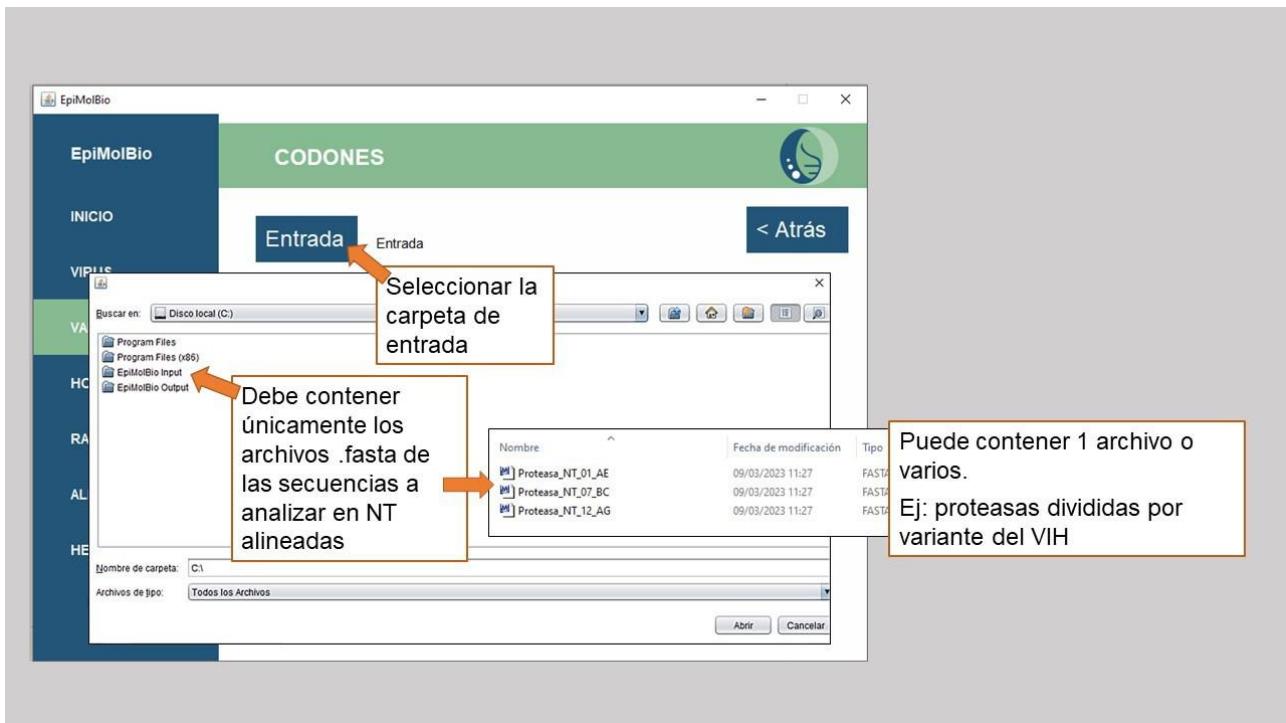
2)



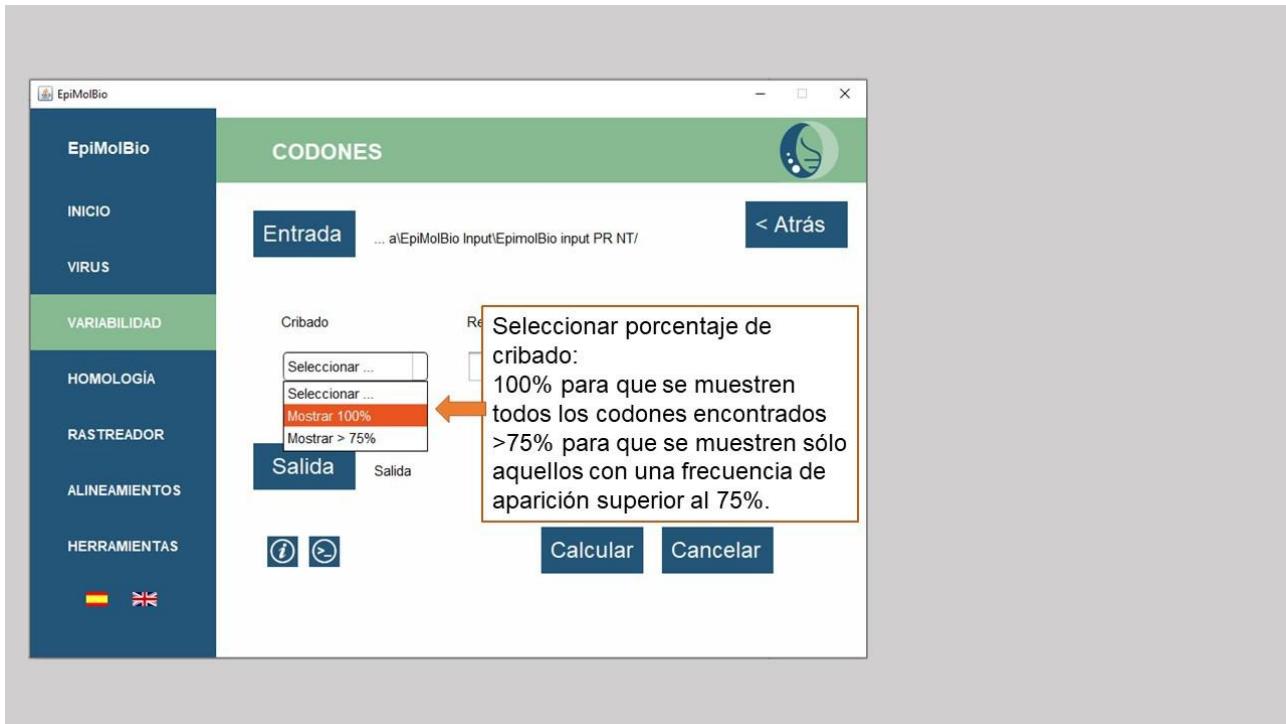
3)



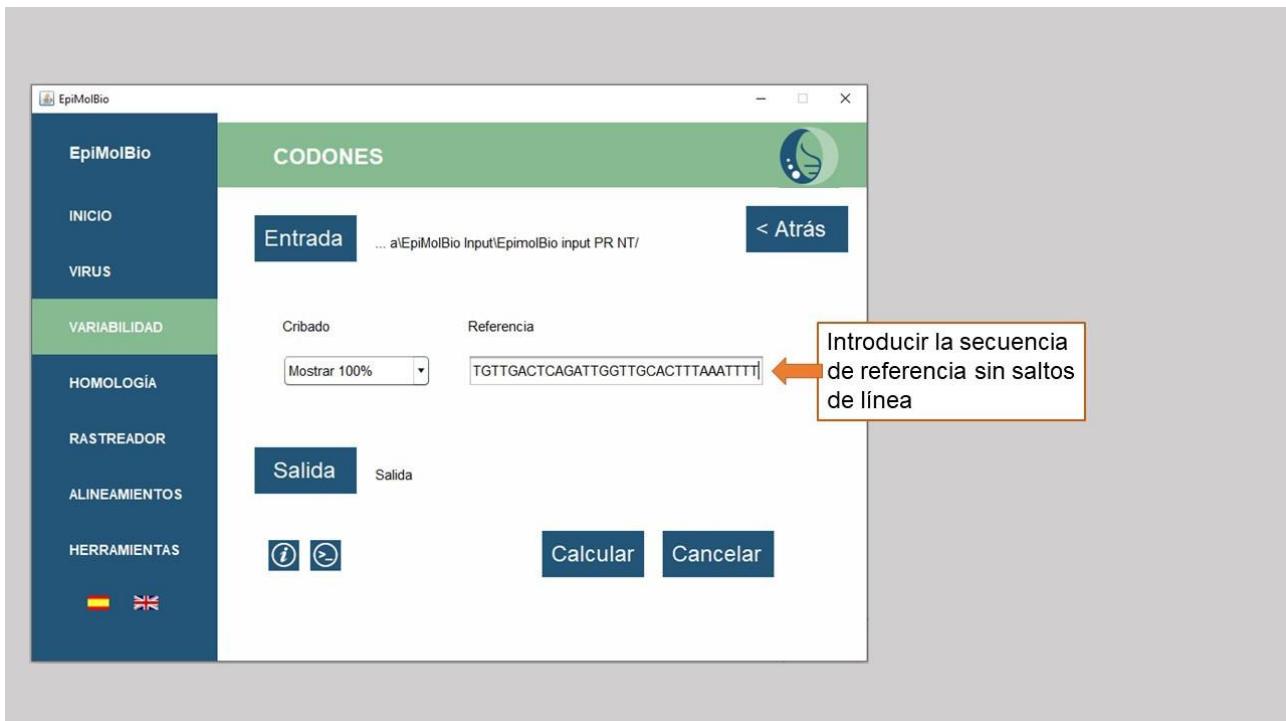
4)



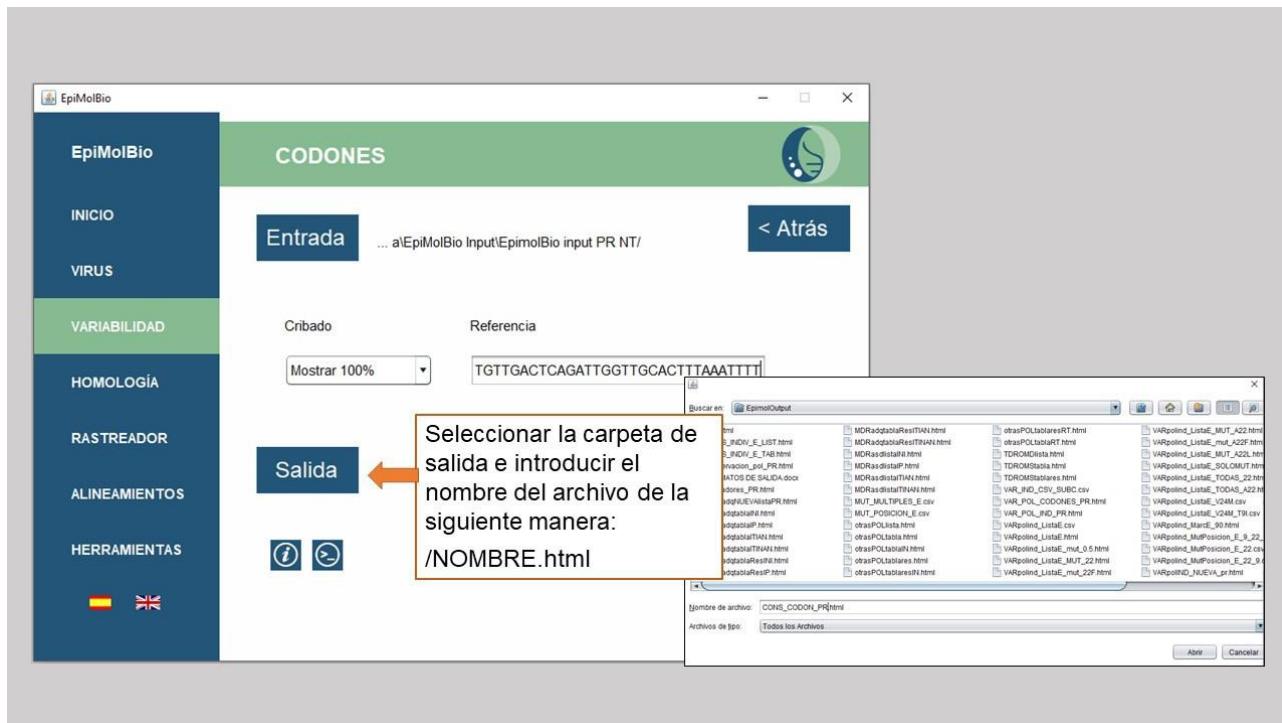
5)



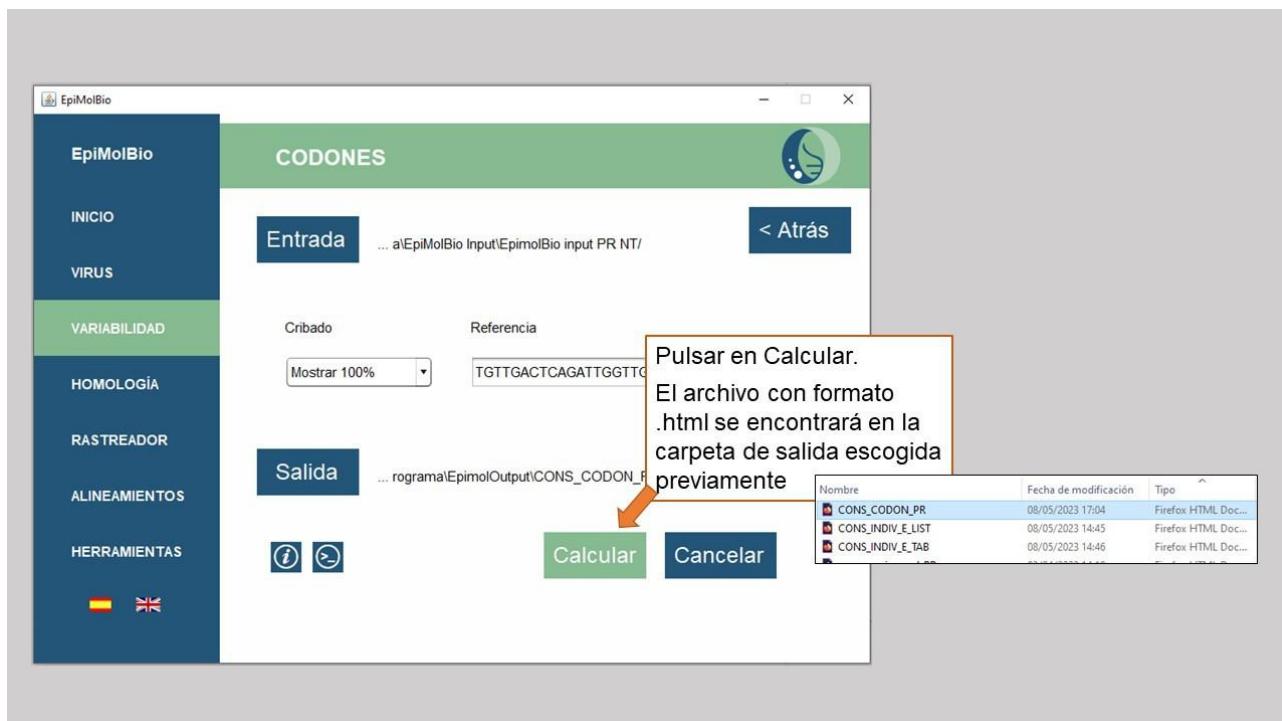
6)



7)



8)



II.3. CONSENSOS

Esta función permite obtener secuencias consenso de otras secuencias consenso generadas a partir de secuencias .fasta introducidas como archivo de entrada. Pueden realizarse varias rondas de análisis sucesivas para obtener los consensos de consensos. Por ejemplo, obtener en una primera ronda las secuencias consenso de distintas variantes de un virus, y en rondas sucesivas la secuencia consenso que englobe los consensos de las variantes previamente procesadas para obtener el consenso de consensos del virus.

Para obtener secuencias de consenso elaboradas a partir de otras secuencias consenso habrá que realizar varios análisis utilizando distintas rondas.

Ronda 1:

El formato del archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar alineadas. Las secuencias pueden ser de nucleótidos o aminoácidos. Para realizar el análisis en secuencias de nucleótidos será necesario emplear la herramienta Buscar y Reemplazar de Edición de Archivos y sustituir las “N” por “?” para que estas se excluyan del análisis.

En el campo “**Seleccionar Ronda**”, seleccionar “**Ronda 1**”.

En el campo “**Referencia**” se debe introducir una secuencia de referencia sin saltos de línea en nucleótidos o aminoácidos según corresponda de acuerdo a los archivos de entrada.

El formato de **salida** de la ronda 1 será un archivo de texto. Habrá que seleccionar la carpeta de salida donde queremos que aparezca sin necesidad de nombrar el archivo. Este archivo de texto servirá como entrada para las rondas sucesivas y se nombra automáticamente “Consensos”. Se recomienda cambiar el archivo de carpeta y renombrarlo antes de repetir este análisis para evitar que se sobrescriba.

Rondas sucesivas:

En **entrada** debemos seleccionar una carpeta donde tengamos **exclusivamente** el **archivo .txt** de la ronda anterior.

En el campo **Seleccionar Ronda**, escoger “**Rondas sucesivas**”.

En el campo “**Referencia**” se debe introducir una secuencia de referencia sin saltos de línea en nucleótidos o aminoácidos según corresponda de acuerdo a los archivos de entrada.

El formato de **salida** será un archivo con extensión .html y otro archivo de texto .txt. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos, sin necesidad de nombrar el archivo. Ambos archivos se nombrarán automáticamente “Consensos”, renombre el archivo antes de repetir este análisis para evitar que se sobrescriba.

Si se quiere hacer más niveles de consensos lo único que hay que hacer es juntar los archivos “.txt” de rondas sucesivas, copiando y pegando en un solo archivo .txt respetando los saltos de línea que estos contienen, y repetir el paso de rondas sucesivas.

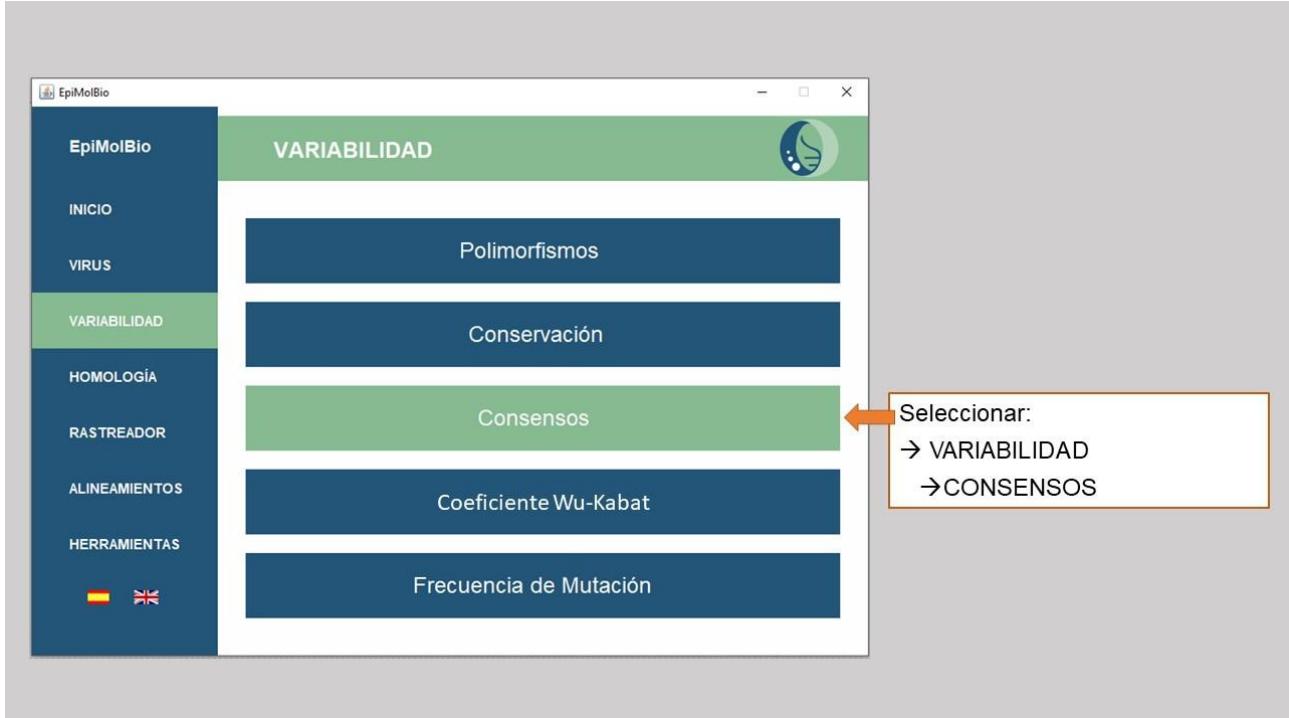
El archivo .html obtenido en rondas sucesivas es una tabla con el consenso obtenido tras el análisis. En la parte superior aparece el título del análisis. En la fila “Referencia” se muestra la secuencia de referencia introducida. En la siguiente fila, “Posición”, la posición del residuo analizado. En la fila “Residuo” aparece el nucleótido o aminoácido más frecuente para cada posición. En la cuarta fila, “Conservación”, aparece el porcentaje de conservación. En la última fila, “Número de Secuencias”, aparece el número de secuencias válidas por cada posición. Las filas 3 y 4 se muestran con las celdas coloreadas según el código de colores descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul. Cabe mencionar que en la tabla de rondas sucesivas no aparece el archivo cargado.

Ejemplo de formato de salida para Rondas Sucesivas de Consensos:

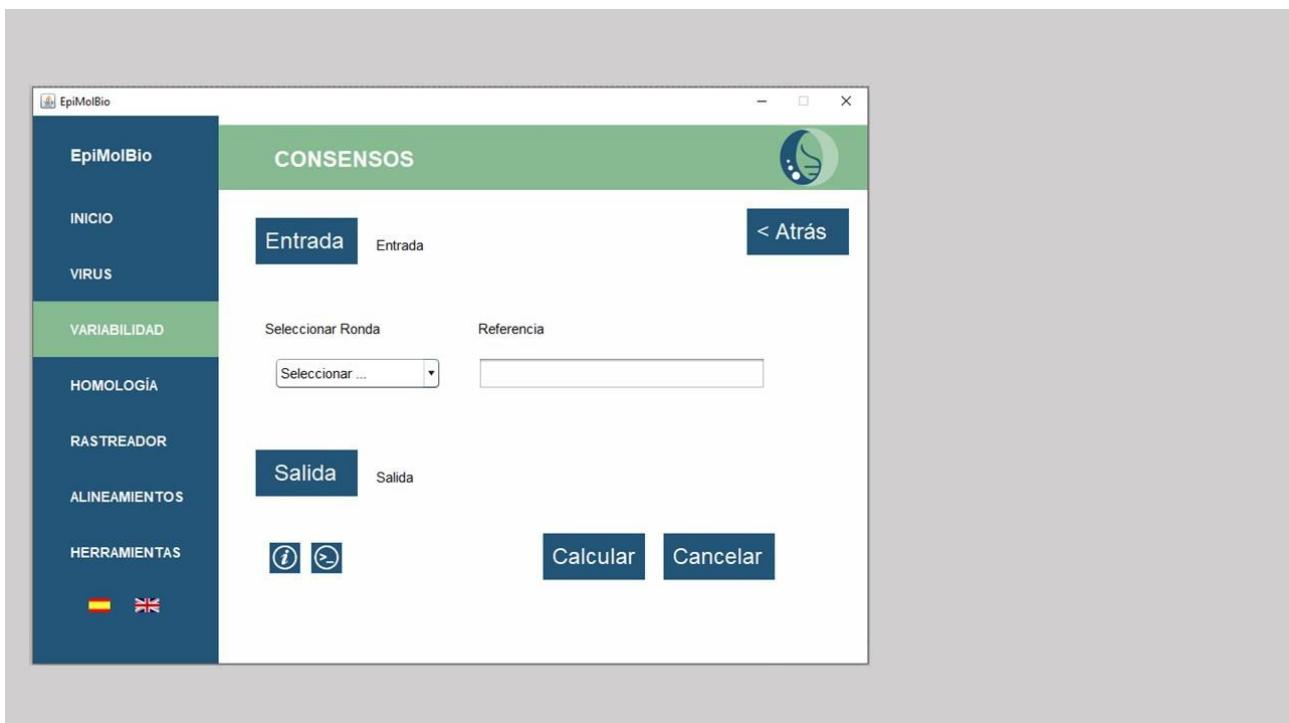
| Variabilidad Consensos | | | | | | | | | | | | | | | | | | |
|------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-----|-----|--|
| Referencia | P | Q | V | T | L | W | Q | R | P | L | V | T | I | K | I | | | |
| Posición | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | | | |
| Residuo | P | Q | I | T | L | W | Q | R | P | L | V | T | I | K | I | | | |
| Conservación | 99.867 | 99.880 | 99.037 | 98.725 | 99.810 | 99.946 | 98.497 | 99.882 | 99.987 | 76.744 | 97.137 | 78.090 | 50.465 | 74.586 | 61.991 | | | |
| Número de Secuencias | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | 146 | |

Paso a paso:

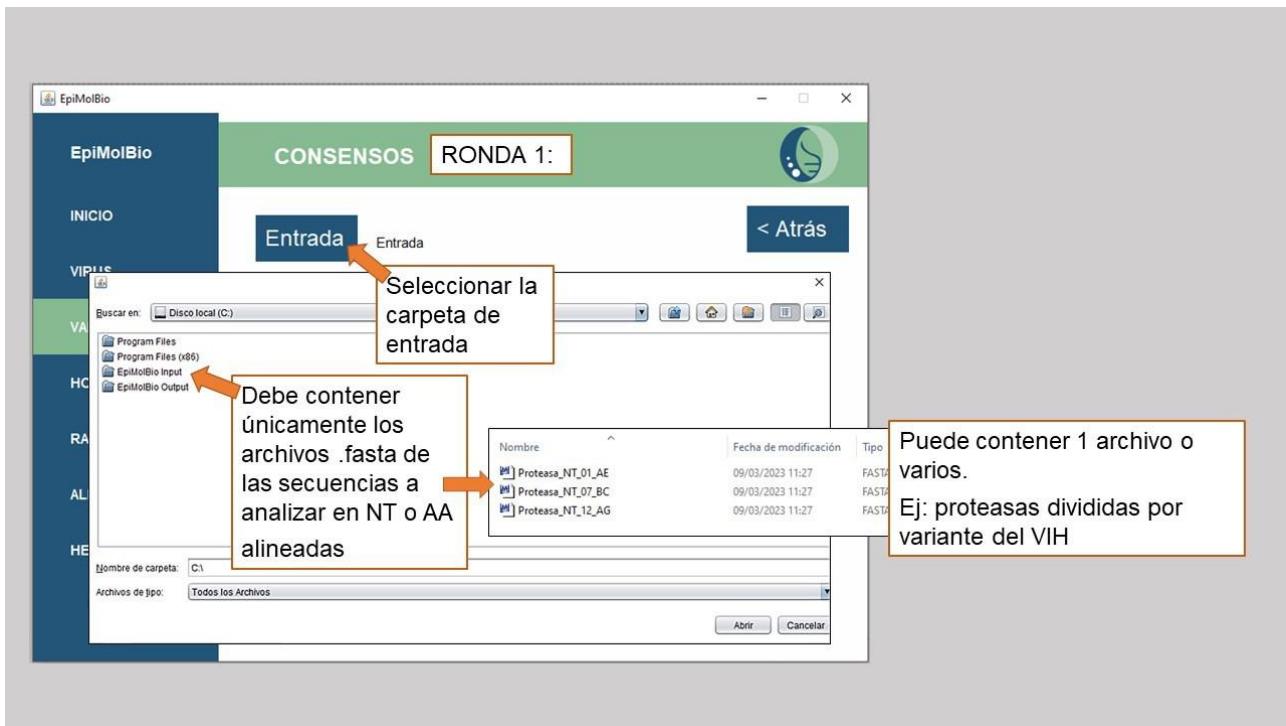
1)



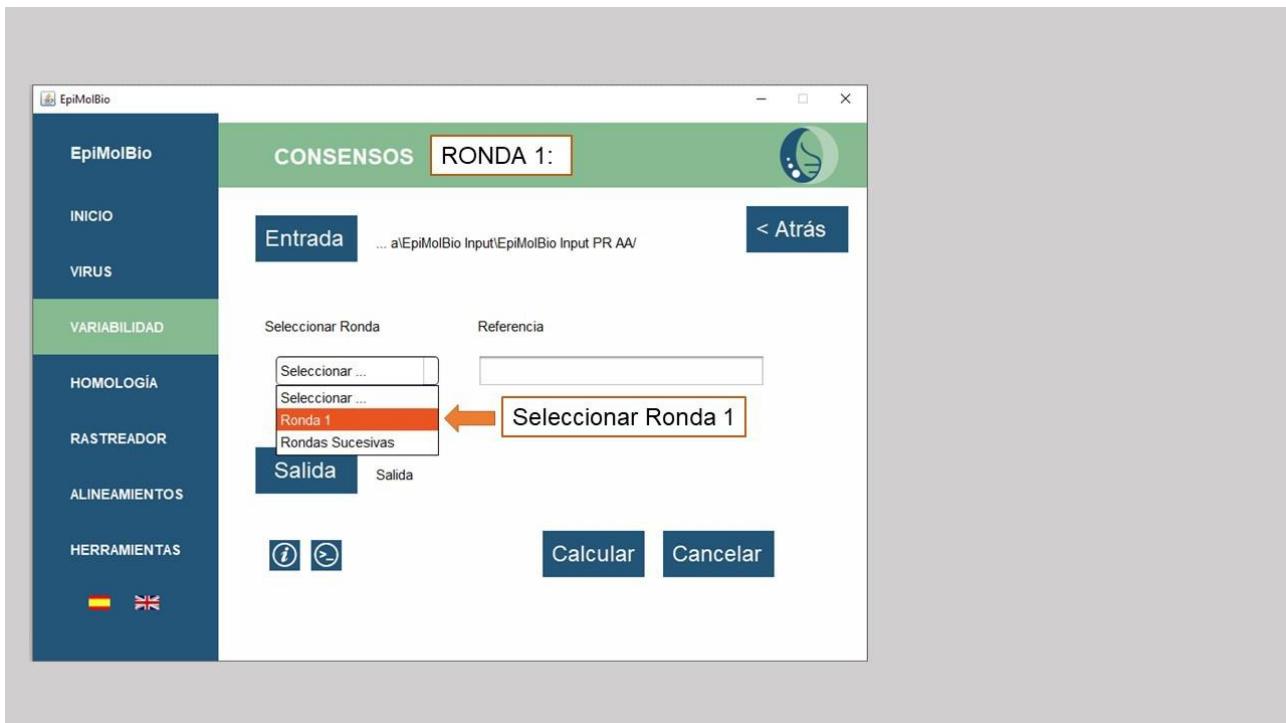
2)



3)



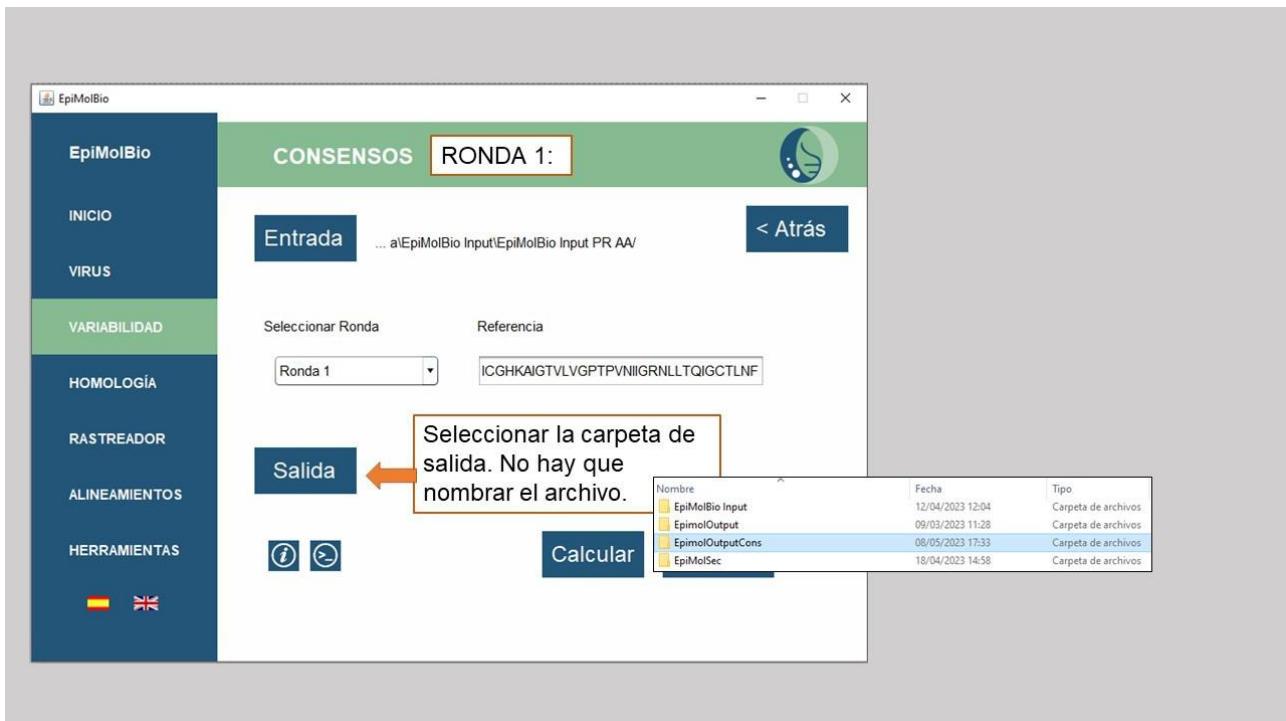
4)



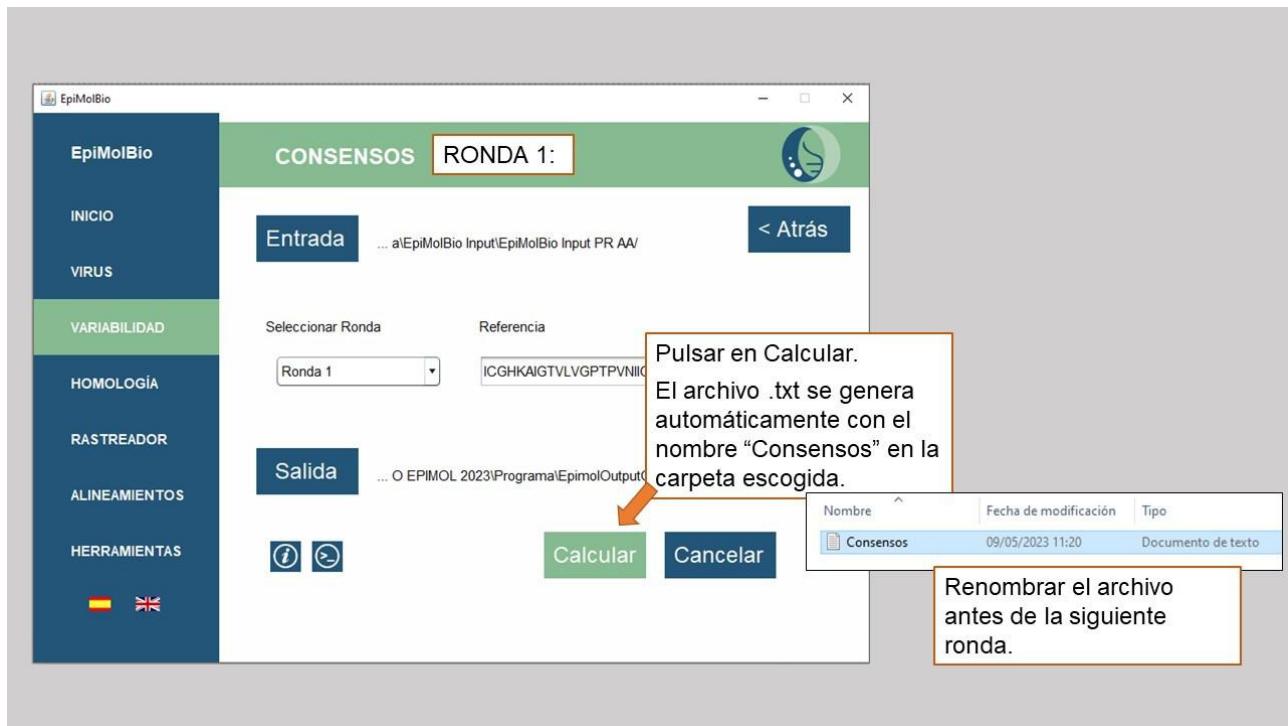
5)



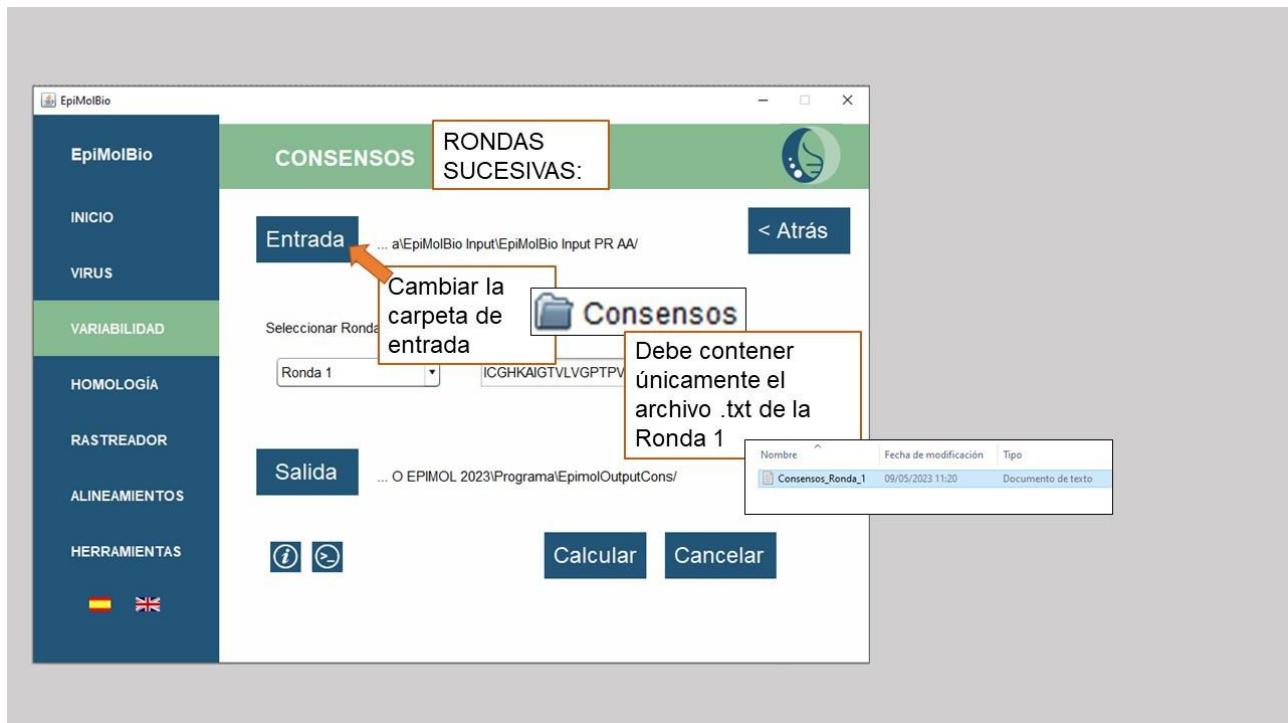
6)



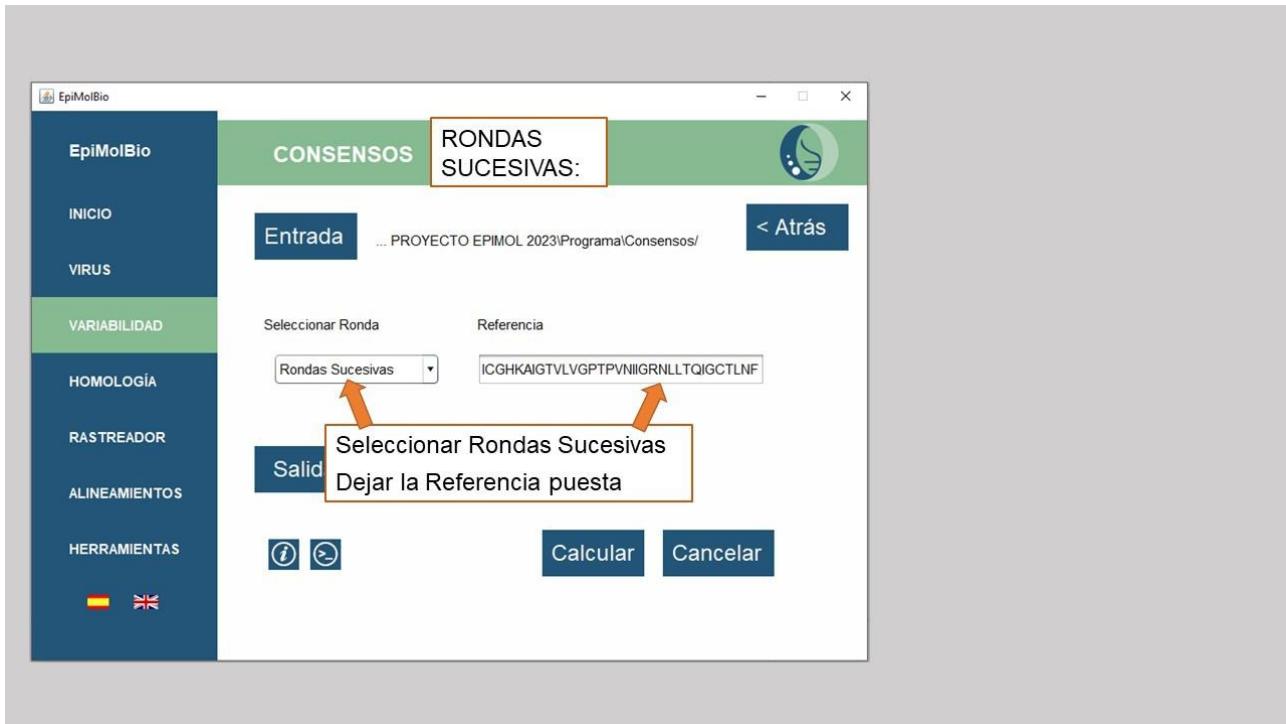
7)



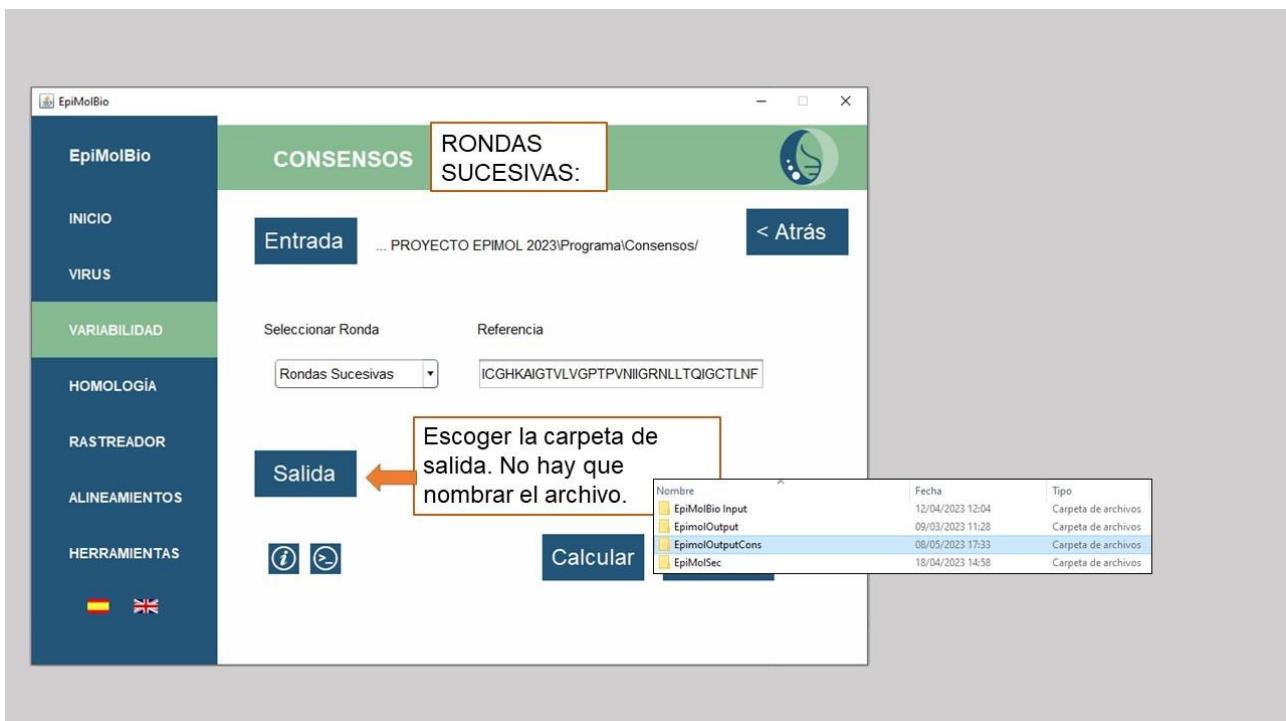
8)



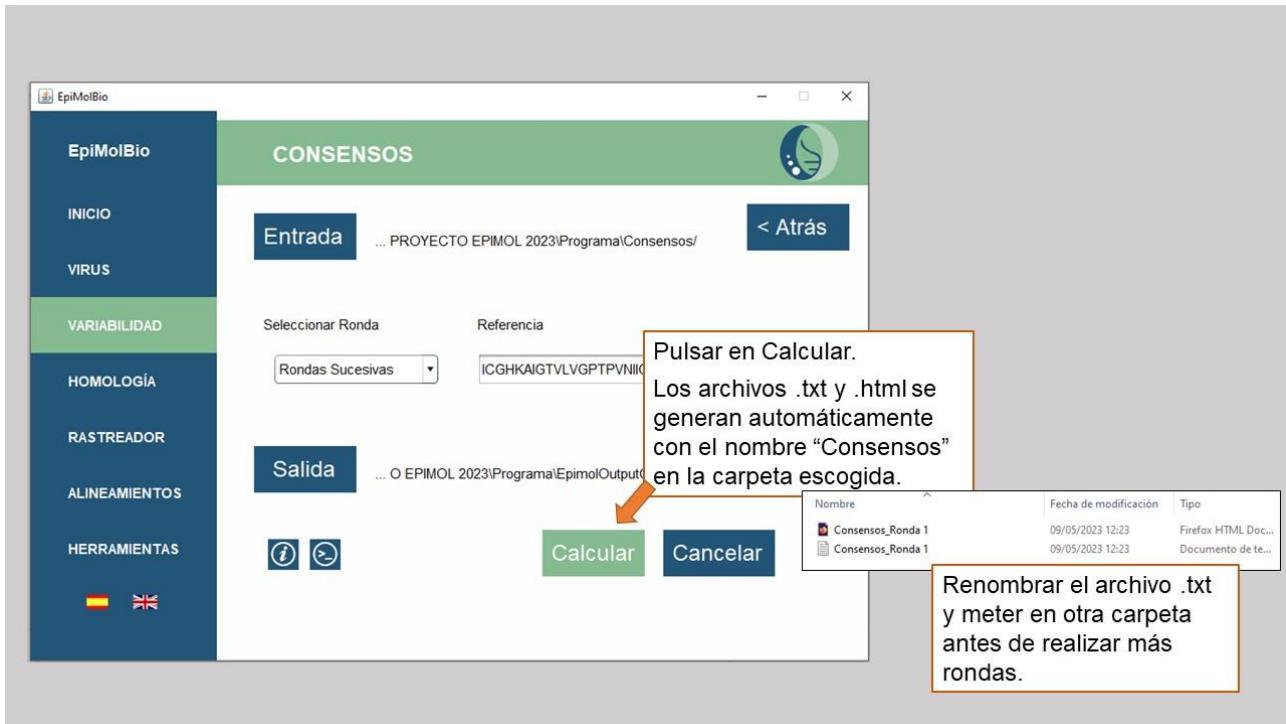
9)



10)



11)



II.4.COEFICIENTE WU-KABAT

Con esta función se puede obtener el coeficiente de variabilidad de Wu-Kabat (WK) de secuencias de proteínas. El coeficiente WK permite estudiar la susceptibilidad de una posición de aminoácidos a los reemplazos evolutivos (Kabat et al., 1977). Se calcula mediante la siguiente fórmula:

$$\text{Variabilidad} = \frac{Nk}{n}$$

Donde N es el número de secuencias en el alineamiento, k es el número de aminoácidos diferentes para una posición determinada y n es el número de veces que está repetido el aminoácido más frecuente en esa posición. Por lo tanto, un WK de 1 indica que se encontró el mismo aminoácido para esa posición en todo el conjunto de secuencias, mientras que un WK >1 indica una variabilidad relativa del sitio respectivo, con una mayor diversidad a medida que aumenta el valor de WK.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar alineadas. Las secuencias deben ser de aminoácidos.

En el campo “**Longitud**” habrá que introducir la longitud en aminoácidos de la proteína a analizar.

El formato de **salida** será un archivo con extensión .csv. Habrá que seleccionar la carpeta de salida donde queremos que aparezca el resultado y nombrar el archivo escribiendo .csv al final.

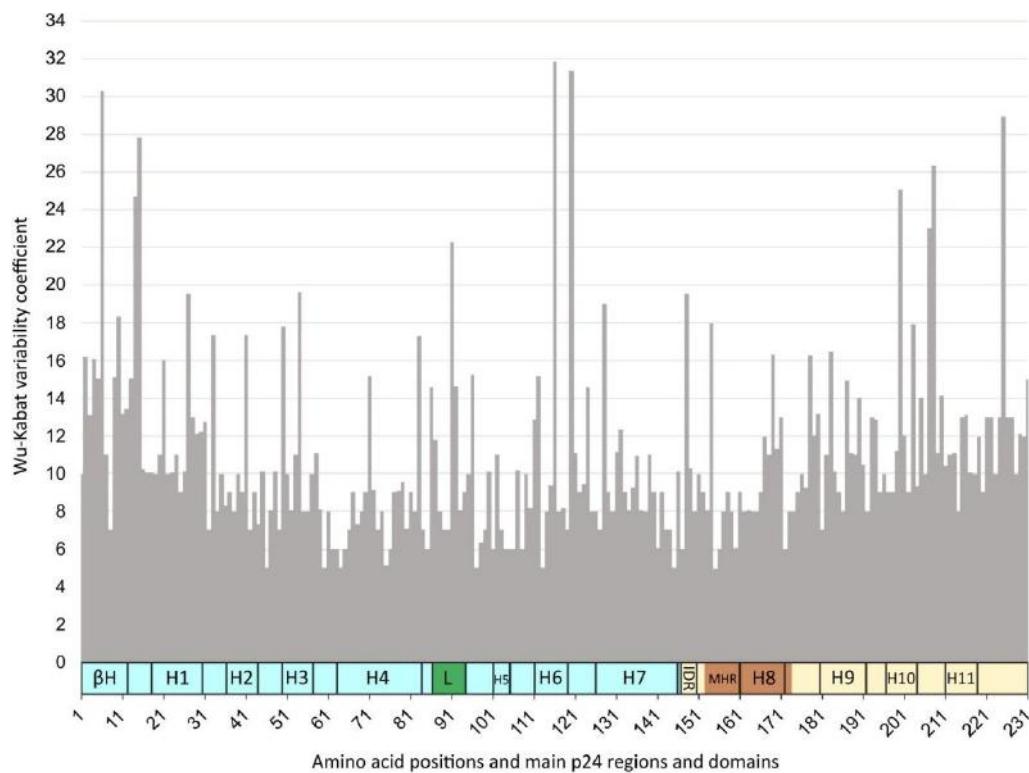
Este formato consiste en una tabla que se puede abrir en Excel. En la tabla se muestra, en la primera columna, el nombre de los archivos de entrada; en la segunda columna, la posición de cada residuo analizado; en la tercera, el índice de Wu-Kabat; en la cuarta, el número de secuencias válidas para esa posición; en la quinta, el número de aminoácidos diferentes en esa posición y en la sexta, la frecuencia absoluta del aminoácido más frecuente para esa posición.

Ejemplo de formato de salida Coeficiente de Wu-Kabat:

| | A | B | C | D | E | F |
|----|----------------|----------|----------|----------------------|-----------------------|------------|
| 1 | Archivo | Posición | Wu-Kabat | Número de Secuencias | Número de Aminoácidos | Frecuencia |
| 2 | PR_01_AE.fasta | 1 | 7.007 | 26838 | 7 | 26810 |
| 3 | PR_01_AE.fasta | 2 | 11.024 | 26649 | 11 | 26591 |
| 4 | PR_01_AE.fasta | 3 | 5.007 | 26831 | 5 | 26793 |
| 5 | PR_01_AE.fasta | 4 | 9.013 | 26816 | 9 | 26778 |
| 6 | PR_01_AE.fasta | 5 | 7.008 | 26780 | 7 | 26750 |
| 7 | PR_01_AE.fasta | 6 | 5.004 | 26836 | 5 | 26817 |
| 8 | PR_01_AE.fasta | 7 | 10.025 | 26536 | 10 | 26470 |
| 9 | PR_01_AE.fasta | 8 | 5.004 | 26792 | 5 | 26771 |
| 10 | PR_01_AE.fasta | 9 | 6.003 | 26613 | 6 | 26601 |
| 11 | PR_01_AE.fasta | 10 | 14.866 | 25952 | 13 | 22694 |
| 12 | PR_01_AE.fasta | 11 | 9.044 | 26416 | 9 | 26287 |
| 13 | PR_01_AE.fasta | 12 | 12.628 | 26469 | 12 | 25153 |
| 14 | PR_01_AE.fasta | 13 | 17.825 | 26150 | 10 | 14670 |
| 15 | PR_01_AE.fasta | 14 | 13.657 | 26270 | 13 | 25006 |

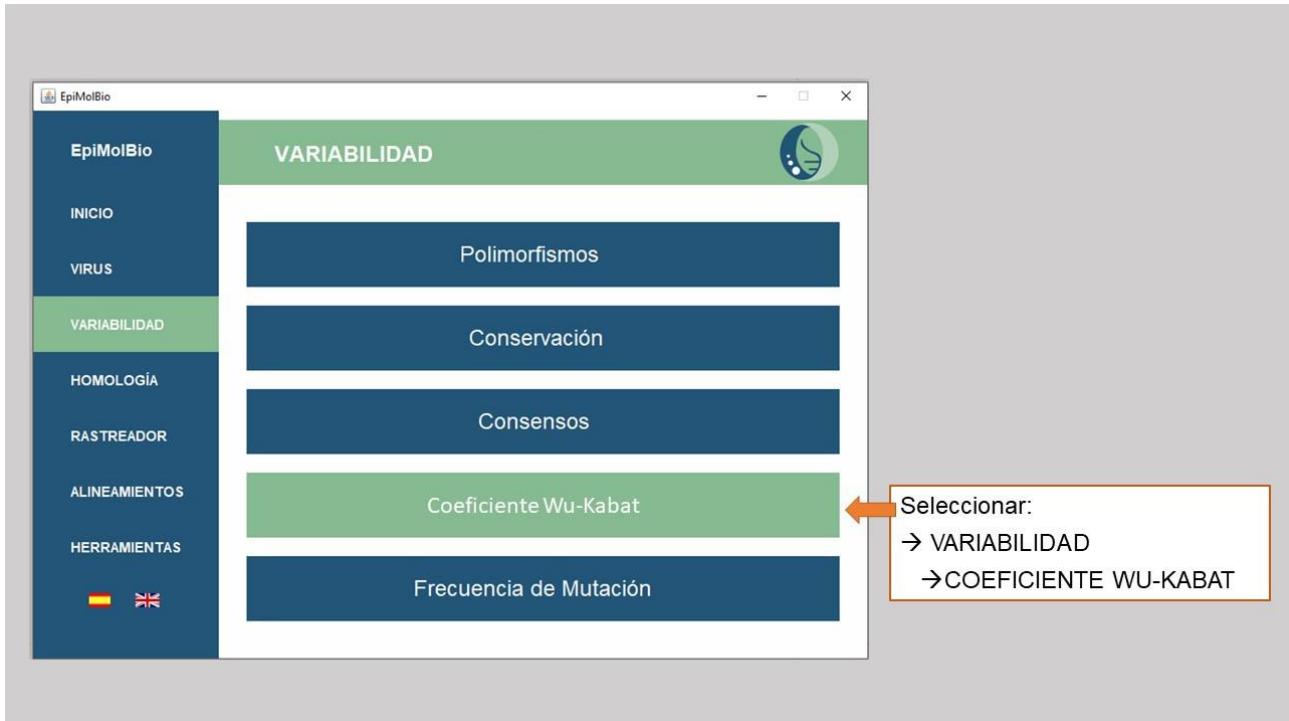
Con las columnas 2 y 3 (Posición y Wu-Kabat) puede elaborarse una gráfica para visualizar el coeficiente de variabilidad de Wu-Kabat de una proteína.

Ejemplo: Diagrama del coeficiente de variabilidad de Wu-Kabat en secuencias de la proteína de la cápside viral p24 del grupo M del VIH-1 (Troyano-Hernáez P, Reinosa R, Holguín Á. HIV Capsid Protein Genetic Diversity Across HIV-1 Variants and Impact on New Capsid-Inhibitor Lenacapavir. *Front Microbiol.* 2022 Apr 12;13:854974. doi: 10.3389/fmicb.2022.854974. PMID: 35495642; PMCID: PMC9039614)

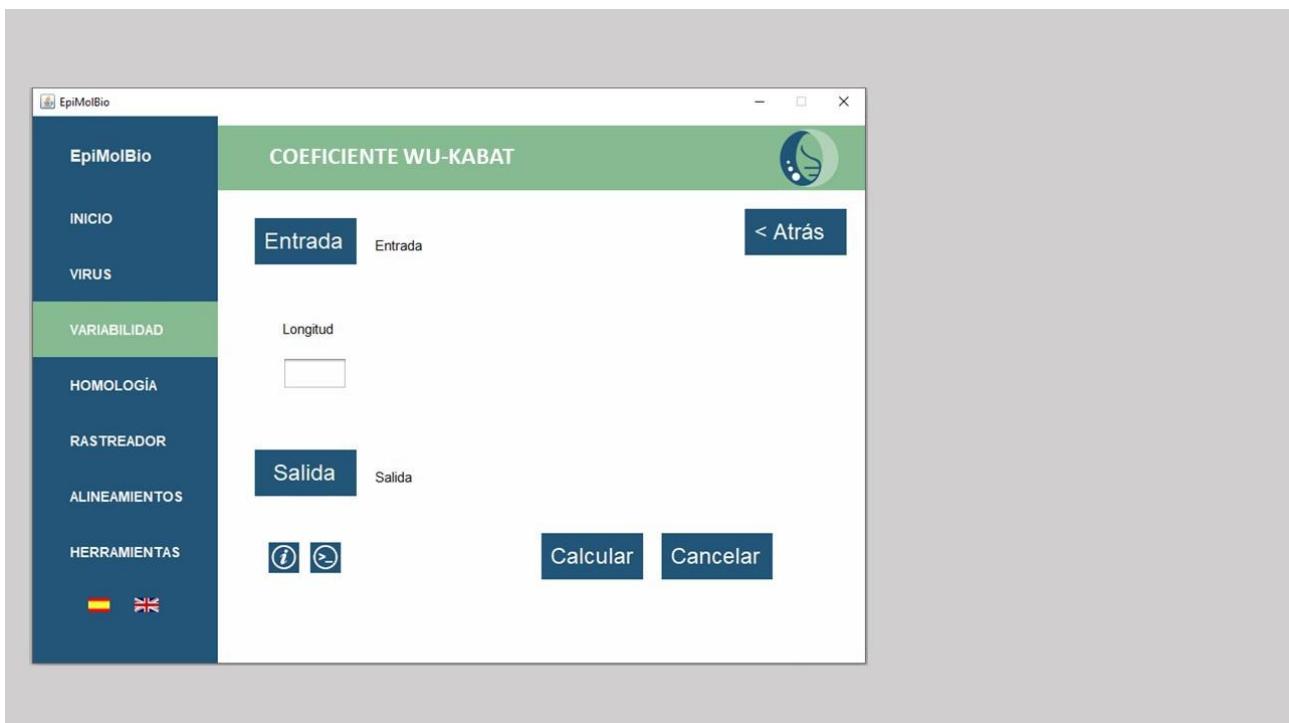


Paso a paso:

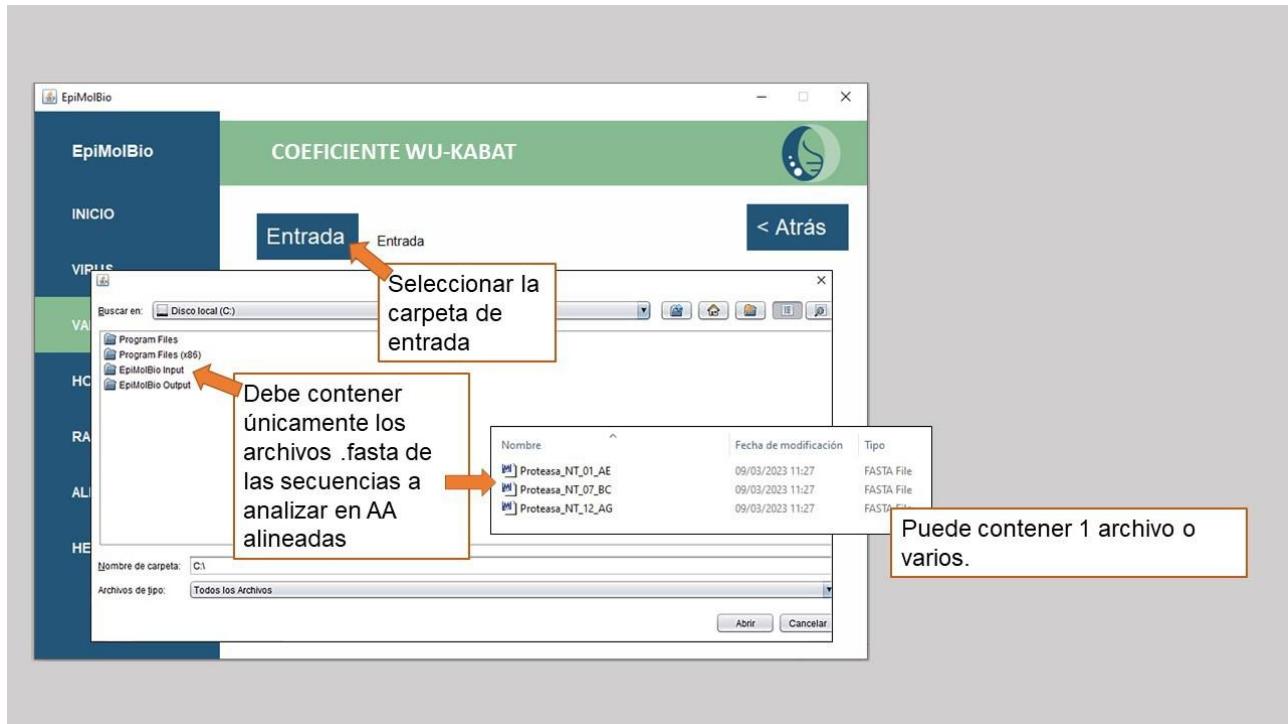
1)



2)



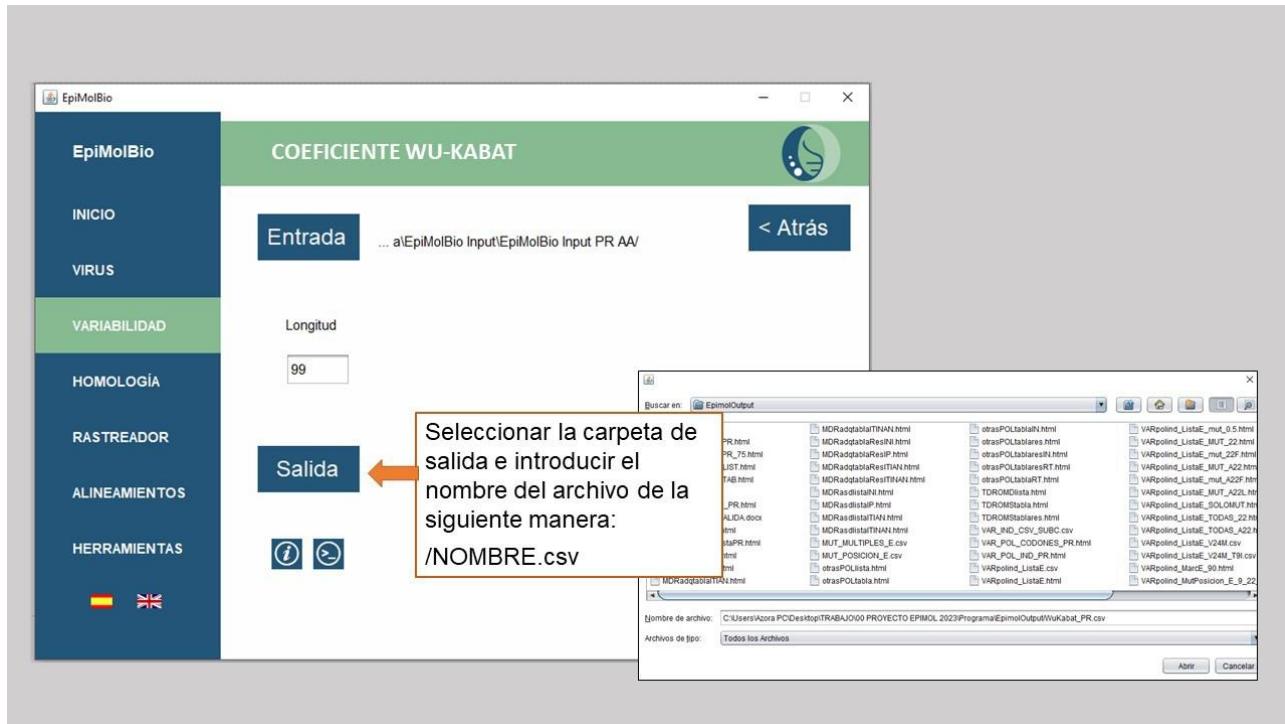
3)



4)



5)



6)



II.5. FRECUENCIA DE MUTACIÓN

Con esta función se pueden obtener una serie de parámetros relacionados con la frecuencia de aparición de mutaciones en un grupo de secuencias de nucleótidos o aminoácidos. Este análisis ignora gaps y residuos faltantes "N" cuando el archivo de entrada es en nucleótidos e ignora gaps, interrogaciones "?" y stops "*", cuando el archivo de entrada está en aminoácidos.

Los parámetros analizados son los siguientes:

Frecuencia de mutación: número de residuos mutados/número de posiciones válidas totales.

Frecuencia de mutación porcentual: frecuencia de mutación x 100.

Porcentaje de conservación: 100 - frecuencia de mutación porcentual.

Mutación media por secuencia: número de residuos mutados / total de secuencias en el archivo de entrada.

El formato del archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar. Las secuencias pueden estar alineadas o sin alinear en el caso de querer detectar inserciones y/o delecciones.

Marcar la casilla "**Nucleótidos**" o "**Aminoácidos**" según si las secuencias del archivo de entrada están traducidas o no.

Marcar el campo "**Alinear**" cuando las secuencias de entrada no estén alineadas. El programa realizará un alineamiento automático, con respecto a la secuencia de referencia, para hacer los cálculos correctamente.

En el campo "**Referencia**" se debe introducir una secuencia de referencia sin saltos de línea y en nucleótidos o aminoácidos según corresponda de acuerdo a los archivos de entrada.

El formato de **salida** será un archivo con extensión .csv. Habrá que seleccionar la carpeta de salida donde queremos que aparezcan el archivo .csv y nombrar los archivos escribiendo .csv al final.

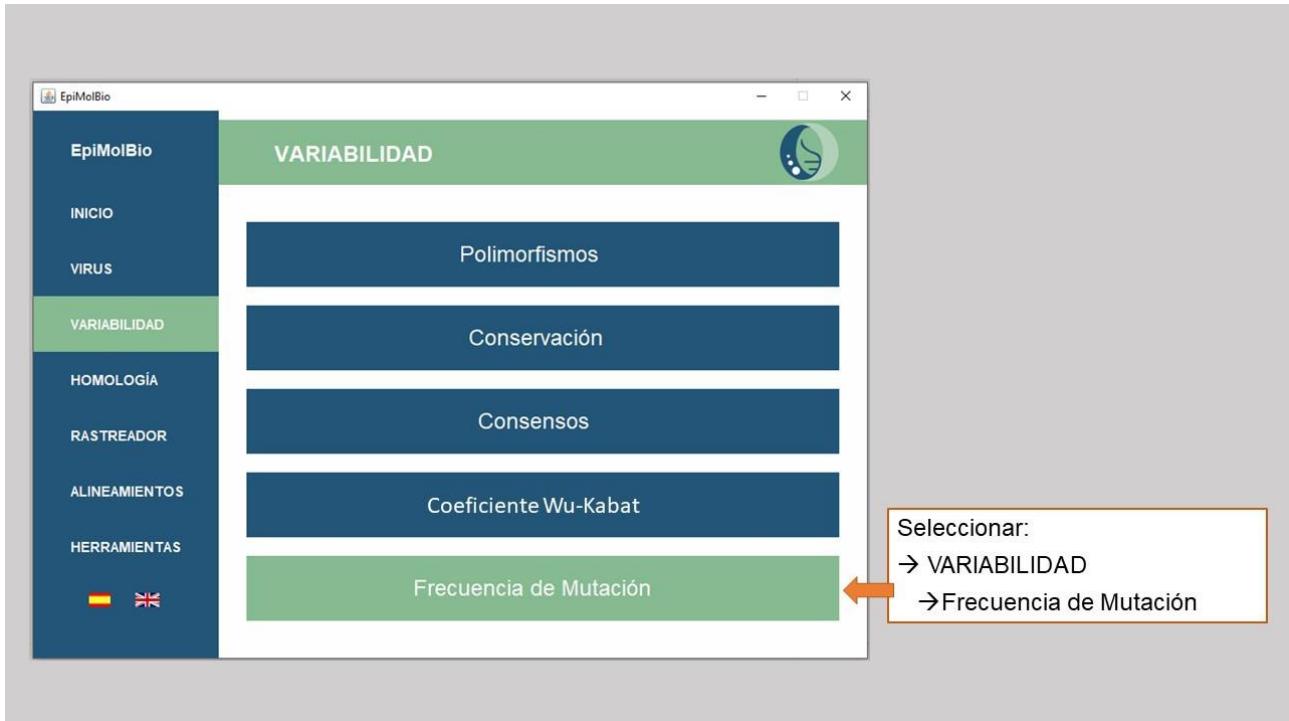
Este formato de salida consiste en una tabla que se puede abrir en Excel. En la tabla se muestra, en la primera columna, el nombre de los archivos de entrada; en la segunda columna, la frecuencia de mutación; en la tercera, la frecuencia de mutación porcentual, en la cuarta; el porcentaje de conservación; en la quinta, el valor de la mutación media por secuencia, y en la última columna, el total de secuencias analizadas por archivo.

Ejemplo de formato de salida de Frecuencia de Mutación:

| | A | B | C | D | E | F |
|---|-------------------------|-----------------|-------------------|----------------|----------------|--------------|
| 1 | Archivo | Frecuencia Mut. | Frecuencia Mut. % | % Conservación | Mut. Media Sec | Sec. Totales |
| 2 | 1-2022_AS_traducido.fas | 0.01333 | 1.33% | 98.67% | 1 | 18 |
| 3 | 1-2022_CL_traducido.fas | 0.01139 | 1.14% | 98.86% | 0.84 | 287 |

Paso a paso:

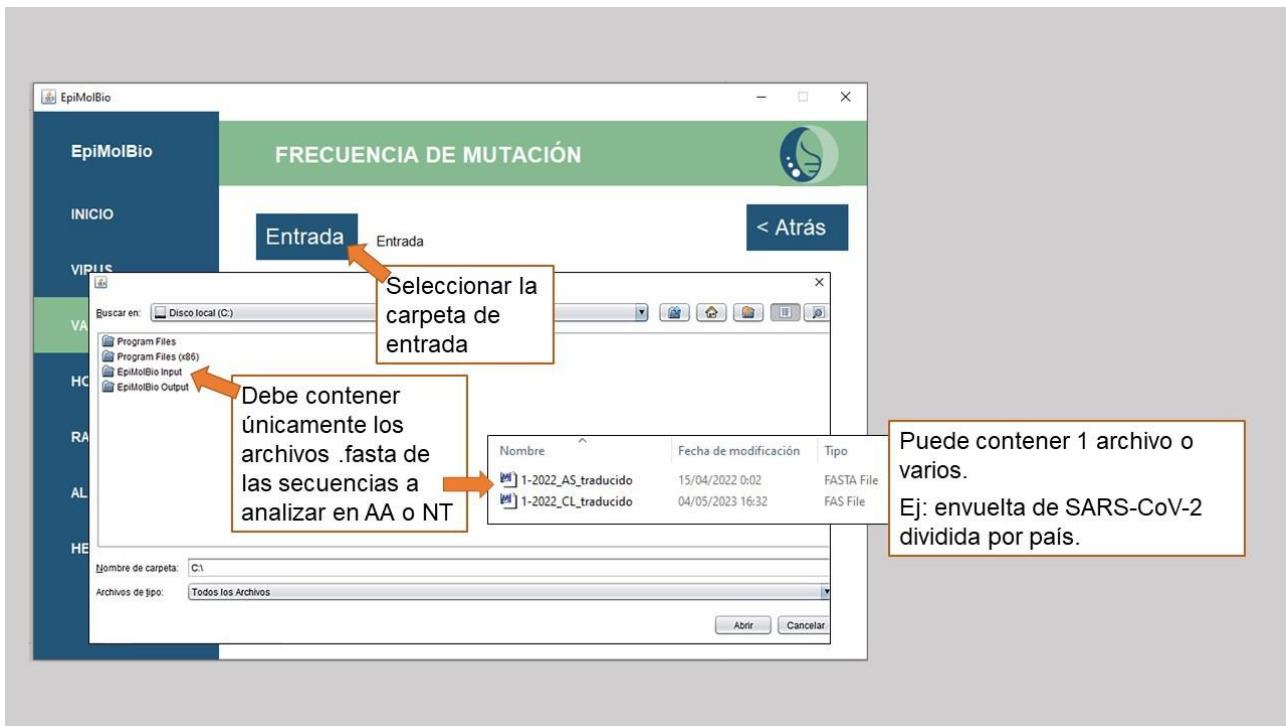
1)



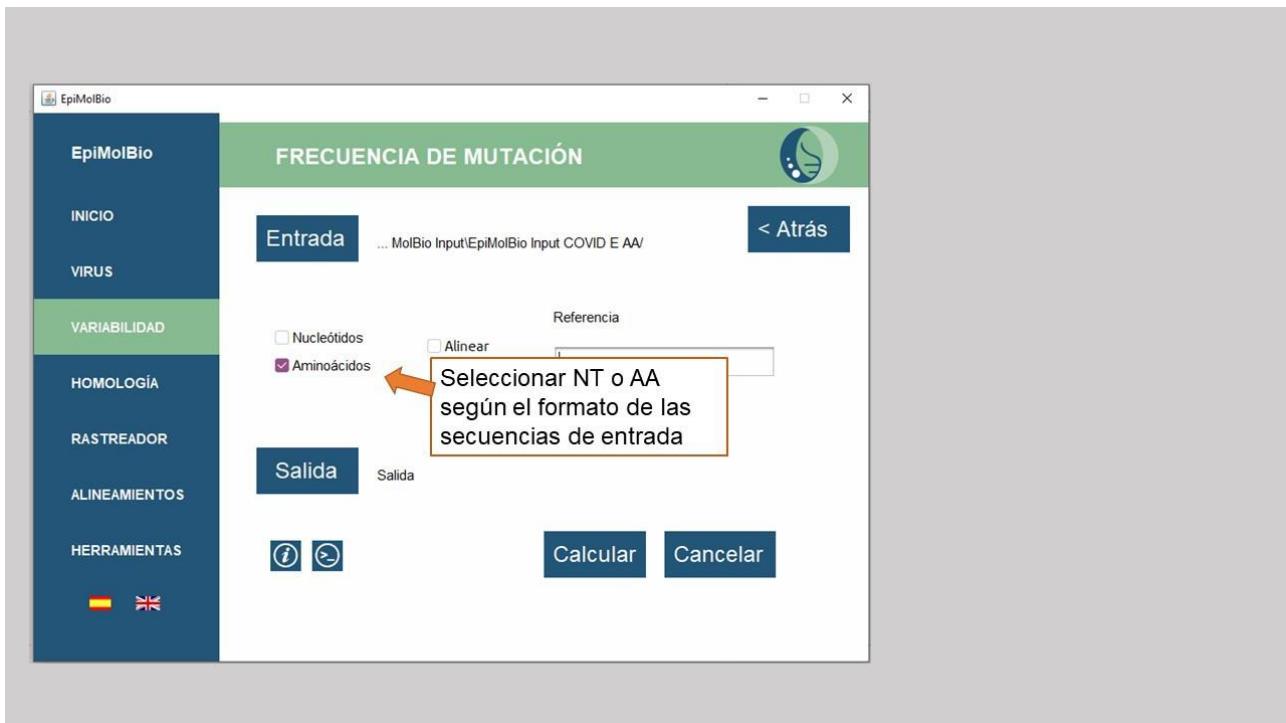
2)



3)



4)



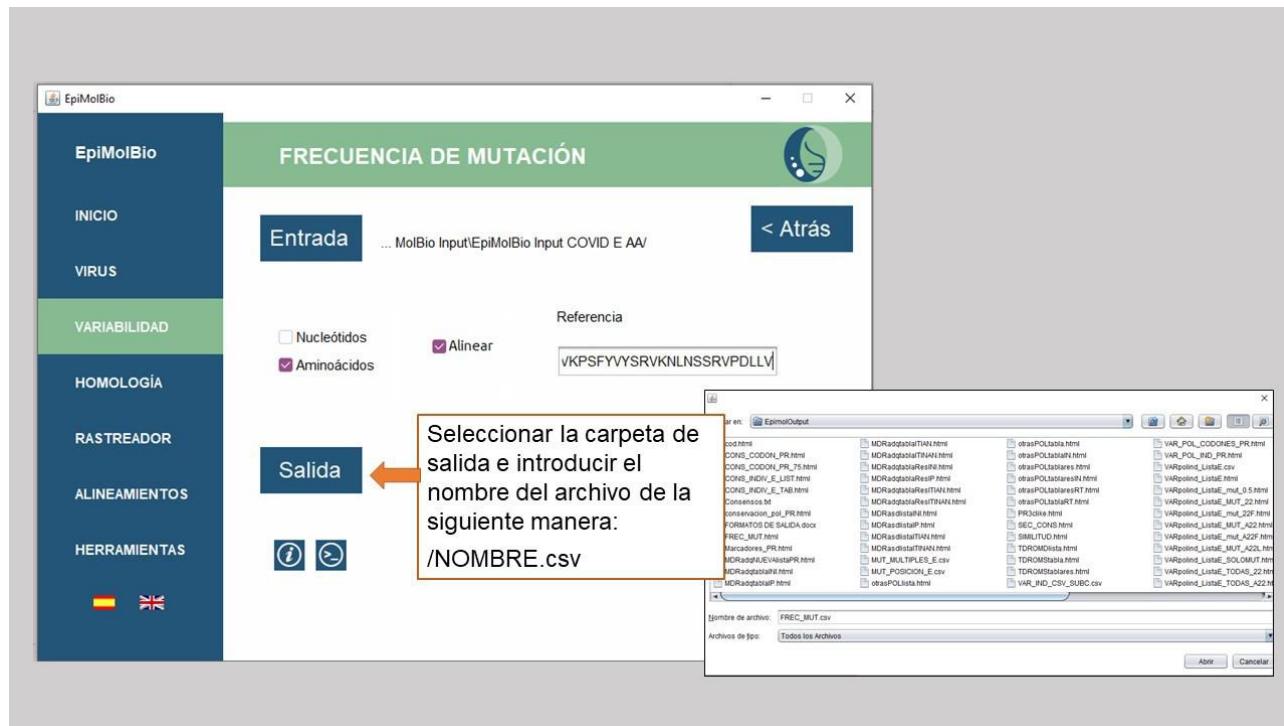
5)



6)



7)



8)



III.HOMOLOGÍA

III.1.SIMILITUD

Esta función permite buscar una secuencia problema concreta introducida por el usuario entre las secuencias del archivo de entrada, obteniendo una tabla .html con la proporción de secuencias por archivo que contienen la secuencia problema. Permite buscar en toda la longitud de la secuencia o en una región concreta. Por ejemplo, se puede utilizar para saber en qué proporción de secuencias se encuentra un péptido concreto.

En el archivo de salida aparece, en la parte superior, el título del análisis seguido de la secuencia problema que se está buscando. Debajo, en la columna “Archivo”, se muestra el nombre de cada archivo analizado. En la columna “Frecuencia”, se muestra la frecuencia de aparición de la secuencia problema en ese archivo, coloreado según el código de colores descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul, y, en la columna “Secuencias Totales”, se muestra el número total de secuencias del archivo analizado.

Ejemplo de formato de salida del análisis Homología Similitud:

| Homología Similitud Rango 5 - 20 | | |
|----------------------------------|------------|--------------------|
| Secuencia Problema: WQRPLVT | | |
| Archivo | Frecuencia | Secuencias Totales |
| PR_01_AE.fasta | 76.643% | 26849 |
| PR_02_AG.fasta | 64.728% | 9577 |
| PR_03_A6B.fasta | 69.355% | 310 |
| PR_04_cpx.fasta | 53.333% | 15 |
| PR_05_DF.fasta | 4.167% | 24 |
| PR_06_cpx.fasta | 61.126% | 746 |
| PR_07_BC.fasta | 76.695% | 10916 |

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias de la proteína a analizar alineadas en nucleótidos o aminoácidos. Las secuencias de entrada pueden no estar alineadas, pero en ese caso no se puede escoger la región de búsqueda en el campo “Rango”.

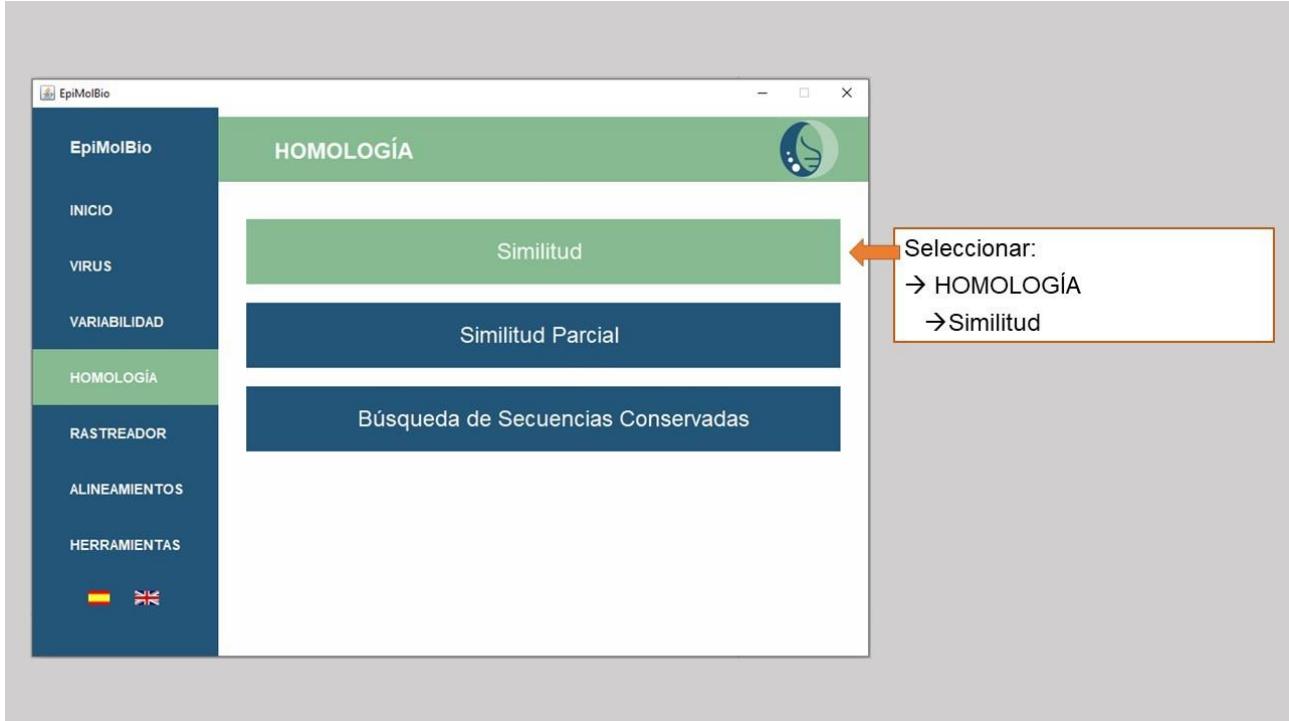
En el Campo “**Secuencia Problema**” escribir la secuencia que se quiere buscar en letras y sin espacios, en nucleótidos o aminoácidos según si los archivos de entrada están traducidos o no (ej.: KLKPGMDGPKVK).

En el Campo “**Rango**” introducir la posición del residuo de la proteína de entrada (en NT o AA) desde donde tiene que comenzar la búsqueda hasta donde tiene que terminarla (ej.: para buscar entre el aminoácido 10 y el 30 inclusive, escribir en la primera caja: “10” y en la segunda, “30”). Si se deja este campo en blanco, buscará en toda la secuencia.

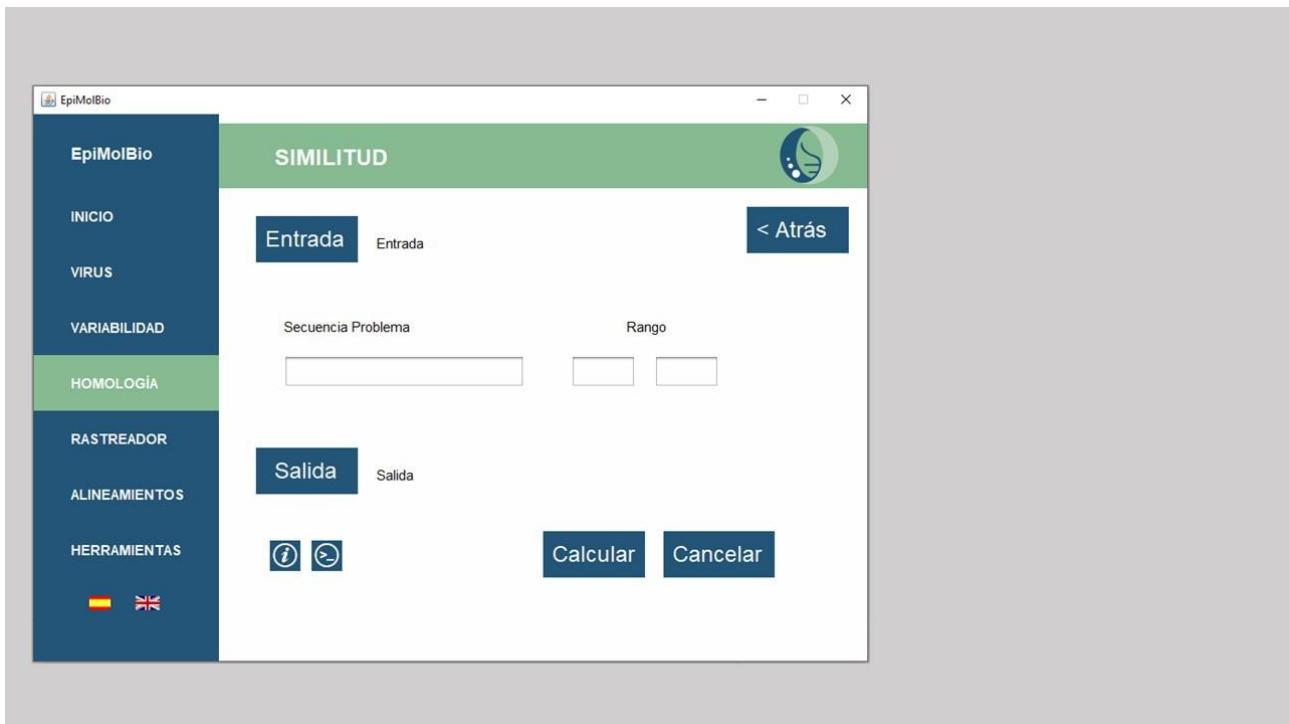
En **salida** habrá que seleccionar la carpeta de salida donde se quiere que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

Paso a paso:

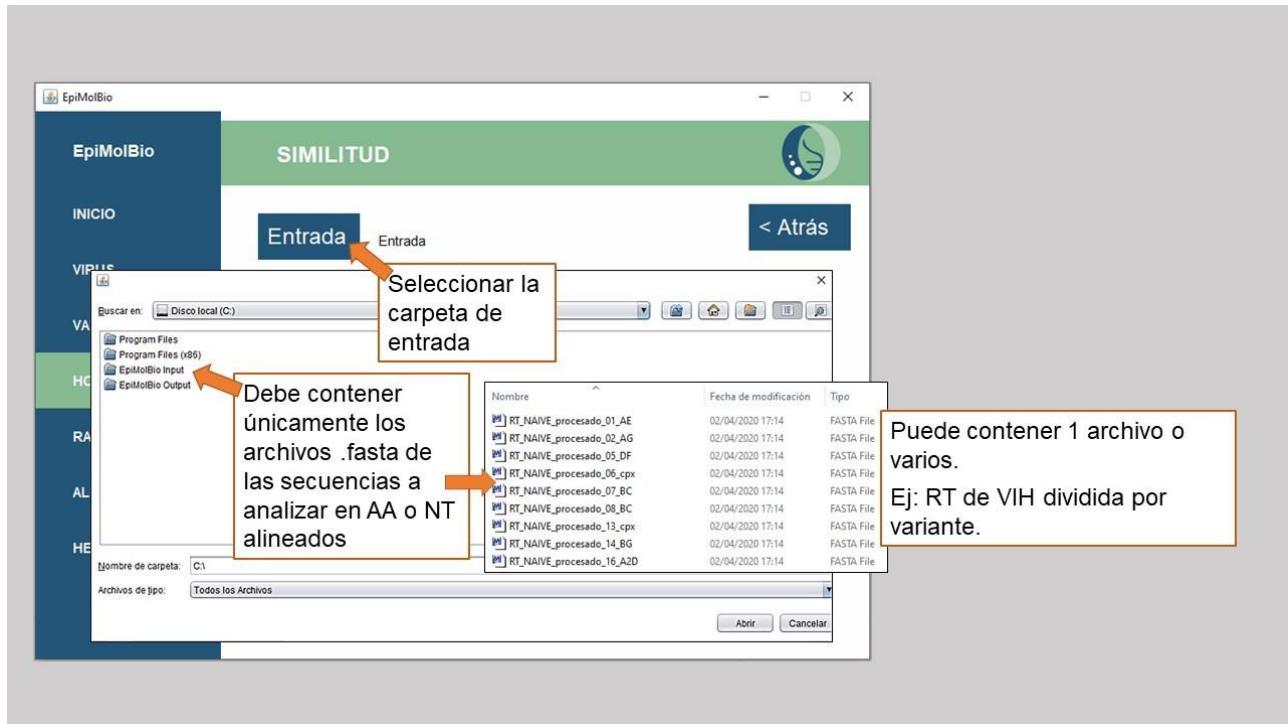
1)



2)



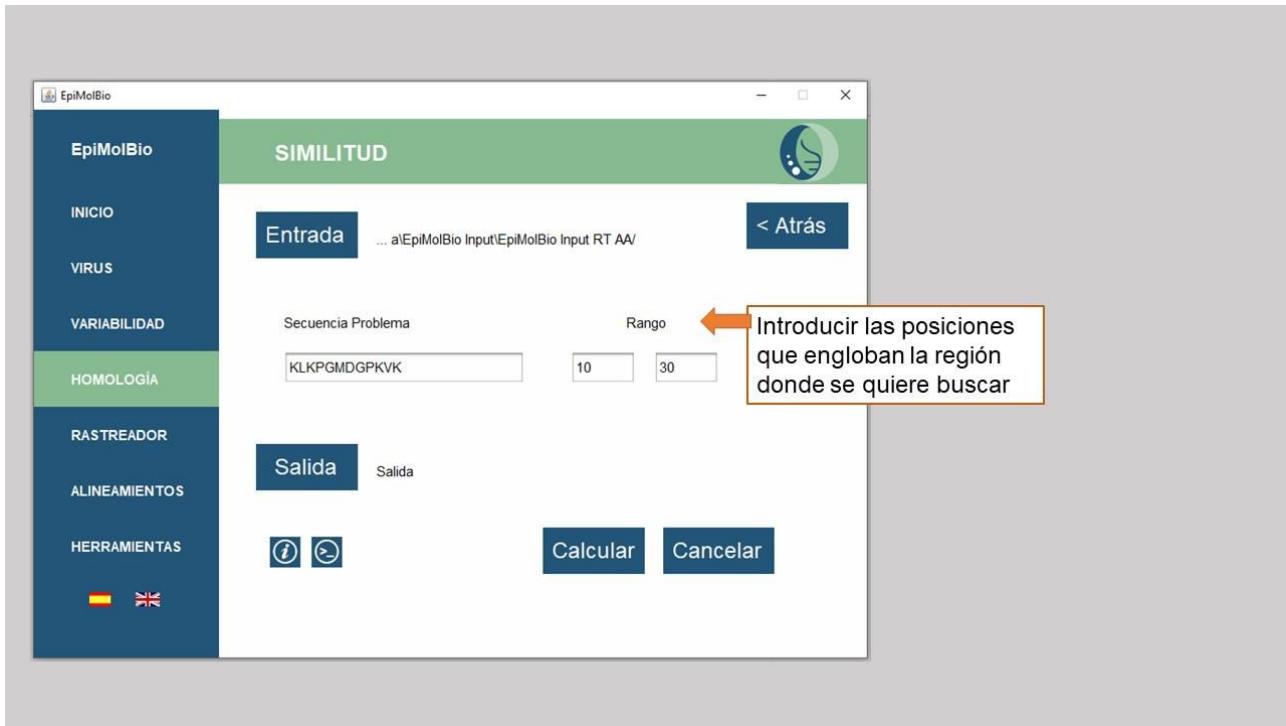
3)



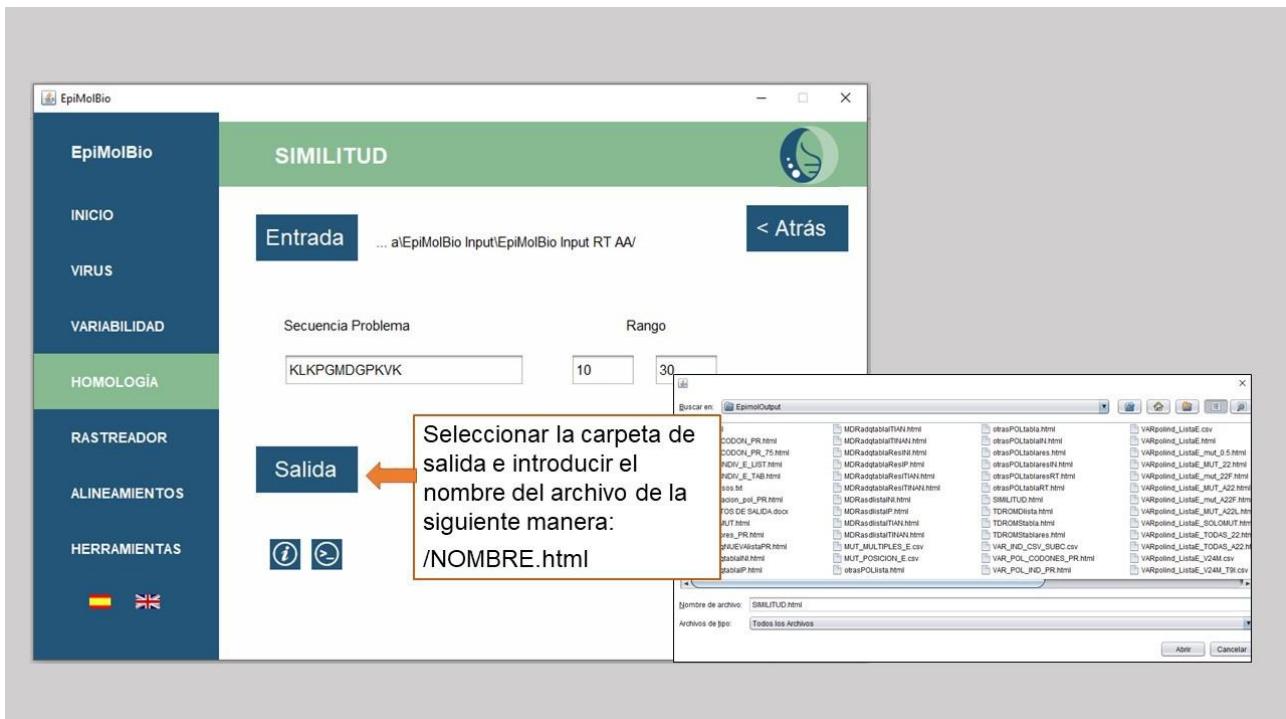
4)



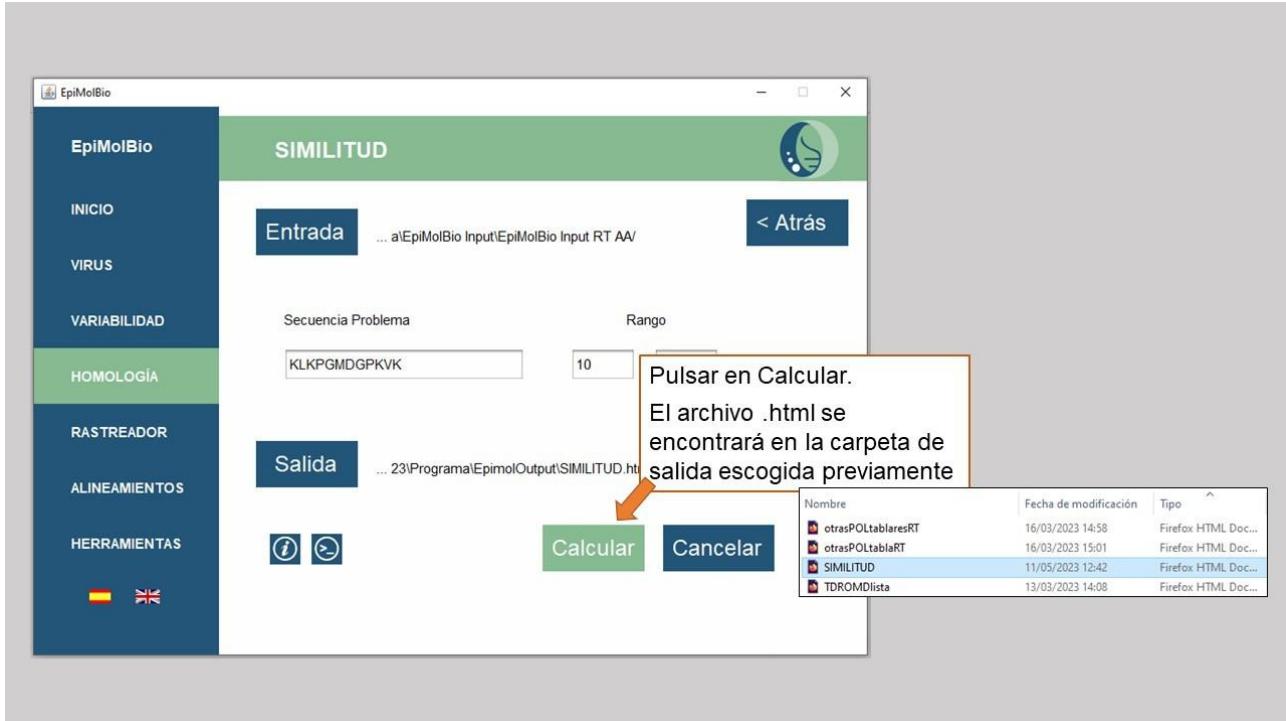
5)



6)



7)



III.2.SIMILITUD PARCIAL

Esta función permite comparar una secuencia introducida por el usuario con las secuencias del archivo de entrada para buscar regiones similares entre ambas, pudiendo definir el porcentaje de similitud. Para ello, el programa corta en varios fragmentos ambas secuencias y los compara, pudiéndose definir la longitud de los fragmentos. Tras el análisis se obtiene una tabla con los fragmentos de secuencias encontrados y su porcentaje de similitud con la secuencia introducida. Por ejemplo: comparar la 3C-like del SARS-CoV-2 (secuencia introducida) con la PR del VIH (archivo de entrada) para localizar regiones de 10 aminoácidos similares en un 50%.

El formato de **salida** es una tabla .html donde aparece, en la parte superior, el título del análisis. Debajo, en la columna “Archivo”, se muestra el nombre del archivo analizado. A continuación, en “Secuencia”, aparece el encabezado de las secuencias del archivo de entrada. En la columna “Secuencia Introducida”, se muestra un fragmento de la secuencia introducida para la búsqueda y en “Secuencia Encontrada”, aparece el resultado obtenido: el fragmento de secuencia encontrada de entre las secuencias del archivo de entrada según los parámetros de análisis escogidos. Junto a ésta, en “Similitud”, aparece el porcentaje de similitud coloreado según el código de colores descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

Ejemplo de formato de salida del análisis Homología Similitud Parcial:

| Homología Similitud Parcial Porcentaje Conservación 50.0% | | | | |
|---|-------------------------------|-----------------------|----------------------|-----------|
| Archivo | Secuencia | Secuencia Introducida | Secuencia Encontrada | Similitud |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | GHRATGTVLV | GTDTTITNVN | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | IGRNLLTQLG | NGMNGRTILG | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | GRNLLTQLGC | GMNGRTILGS | 50.000% |
| PR_D.fasta | >D.ET.2003.ETH_G_230.AB285830 | NLLTQLGCTL | NGRTILGSAL | 50.000% |
| PR_D.fasta | >D.JP-.patient_88.AB356098 | YDQIHVEICG | LTQDHVDILG | 50.000% |
| PR_D.fasta | >D.JP-.patient_88.AB356098 | DQIHVEICGH | TQDHVDILGP | 50.000% |
| PR_D.fasta | >D.JP-.patient_88.AB356098 | QIHVEICGHK | QDHVDILGPL | 50.000% |
| PR_D.fasta | >D.SN.1990.SE365.AB485648 | DDTVLEDINL | FTTTLNDFNL | 50.000% |

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias de la proteína a analizar en nucleótidos o aminoácidos. Se recomienda ajustar el número de secuencias según la longitud de estas para no alargar la velocidad de procesamiento en exceso. Si las secuencias son muy largas (ej.: genoma completo), es preferible realizar búsquedas por tandas.

En el campo “**Longitud**” introducir la longitud de los fragmentos en que se divide la secuencia introducida y la secuencia de entrada para compararlas (ej.: 10, las secuencias se dividirán en trozos de 10 aminoácidos o nucleótidos para compararse).

En el campo “**% Similitud**” definir el porcentaje de similitud mínimo que queremos que tengan los fragmentos comparados para que aparezcan en los resultados, introduciéndolo en cifras con un decimal sin el símbolo de “%” (ej.: 50.0 para una similitud del 50%).

En el campo “**Alinear**”, si se escoge “Alinear”, los fragmentos se alinean entre sí realizando una búsqueda más exhaustiva pero más lenta. Se puede escoger “No Alinear” para una búsqueda más rápida pero menos profunda. Se recomienda emplear la segunda

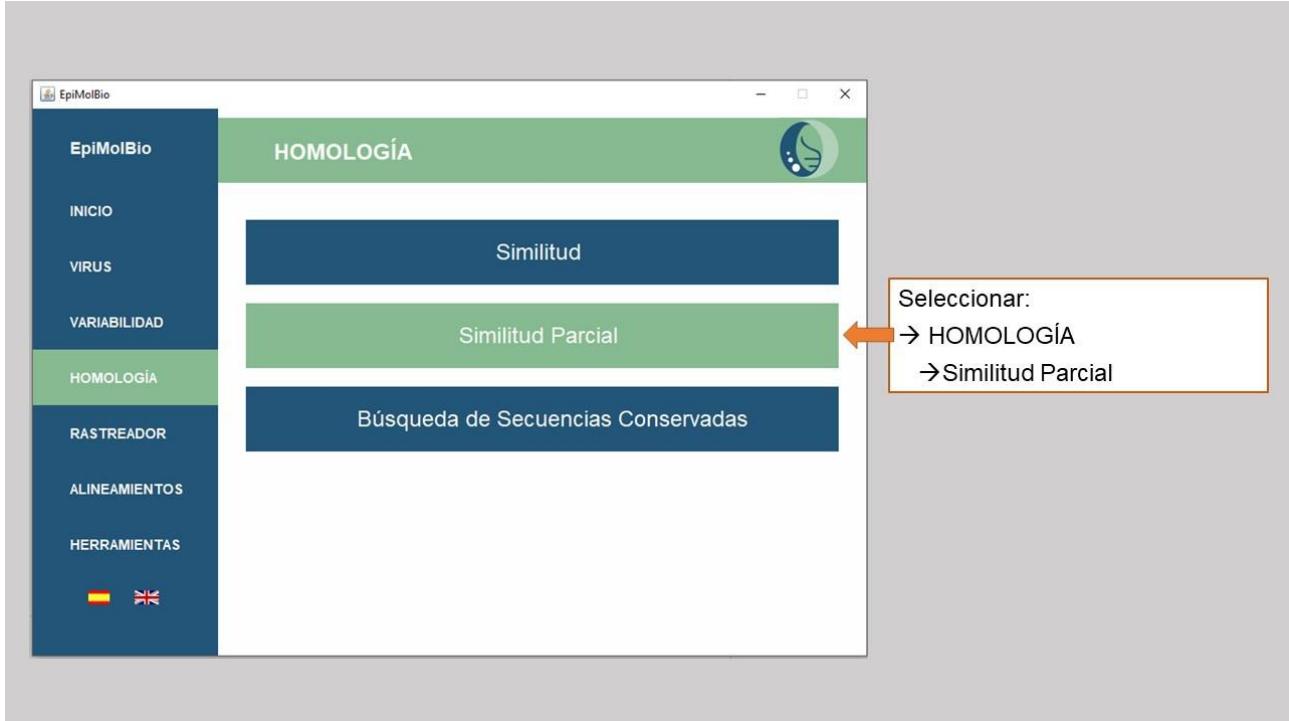
opción para un primer vistazo y la primera para el análisis exhaustivo en caso de haber encontrado resultados relevantes.

En el Campo “**Secuencia**” escribir la secuencia que se quiere buscar en nucleótidos o aminoácidos según el archivo de entrada. La secuencia no puede contener saltos de línea ni espacios (continuando con el ejemplo anterior: si la entrada son secuencias de la PR del VIH, en Secuencia introducir la secuencia completa de la 3C-like del SARS-CoV-2).

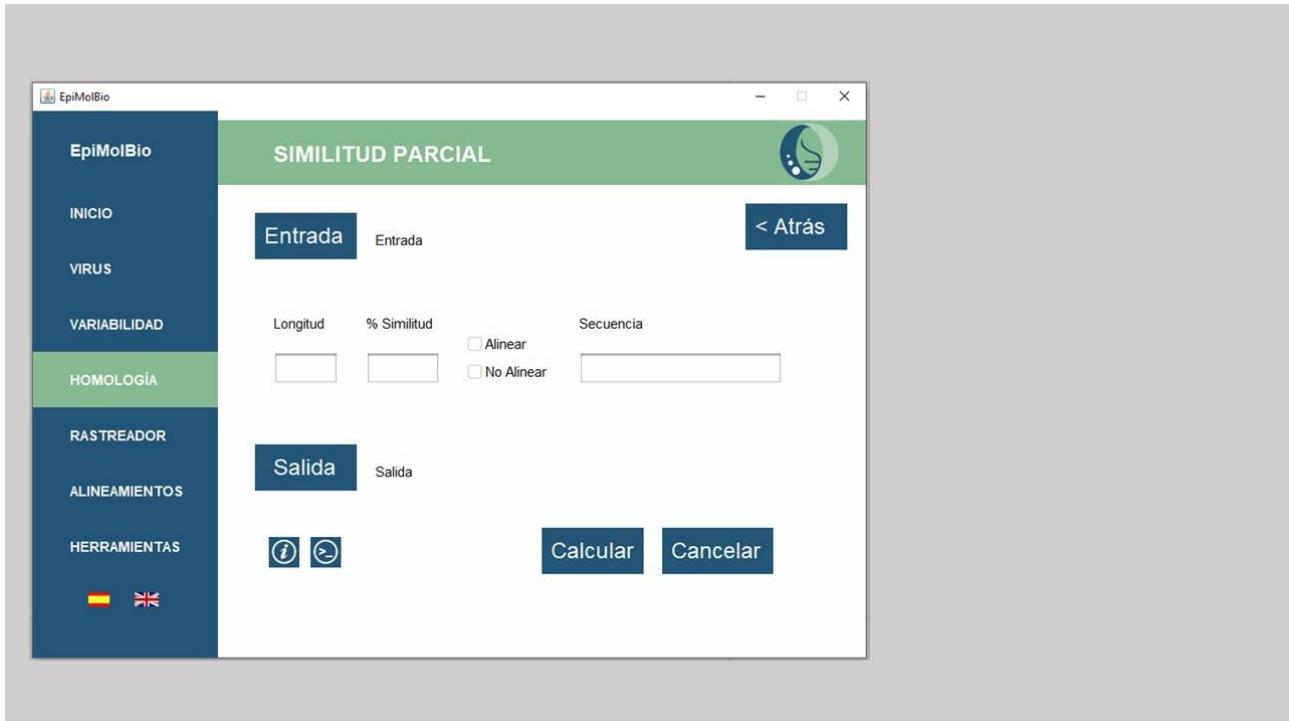
En **salida** habrá que seleccionar la carpeta de salida donde se quiere que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

Paso a paso:

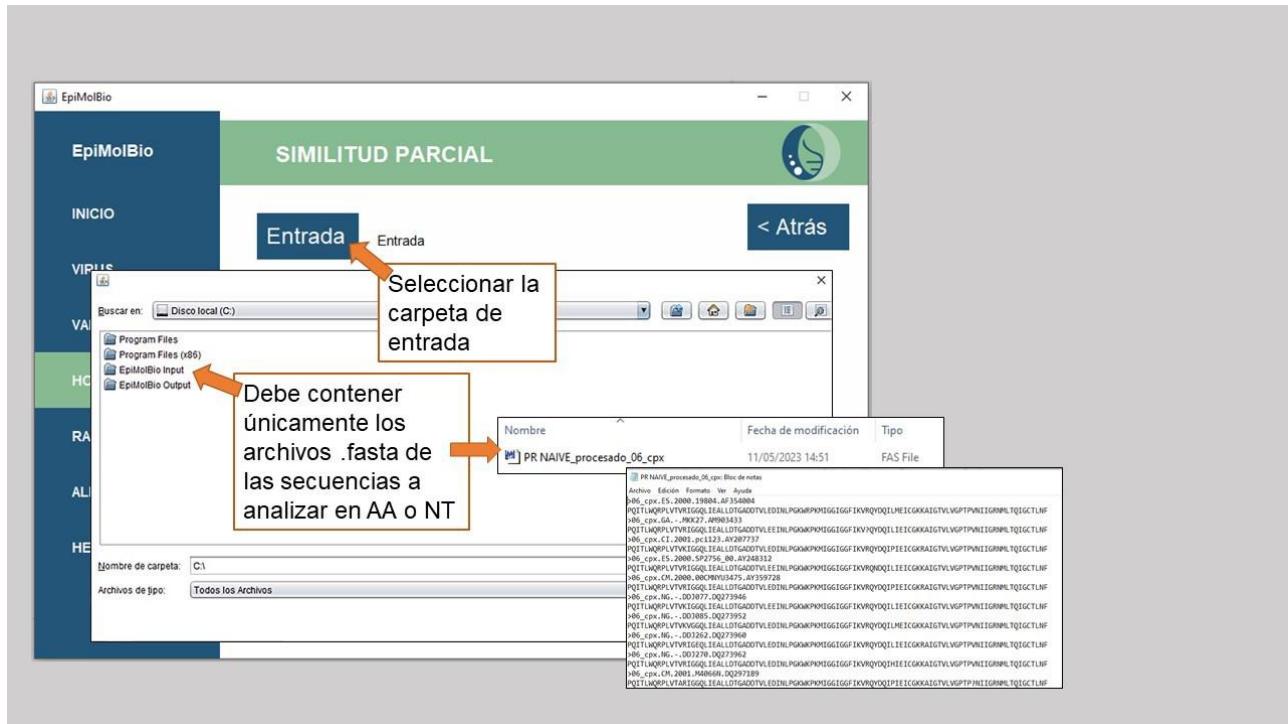
1)



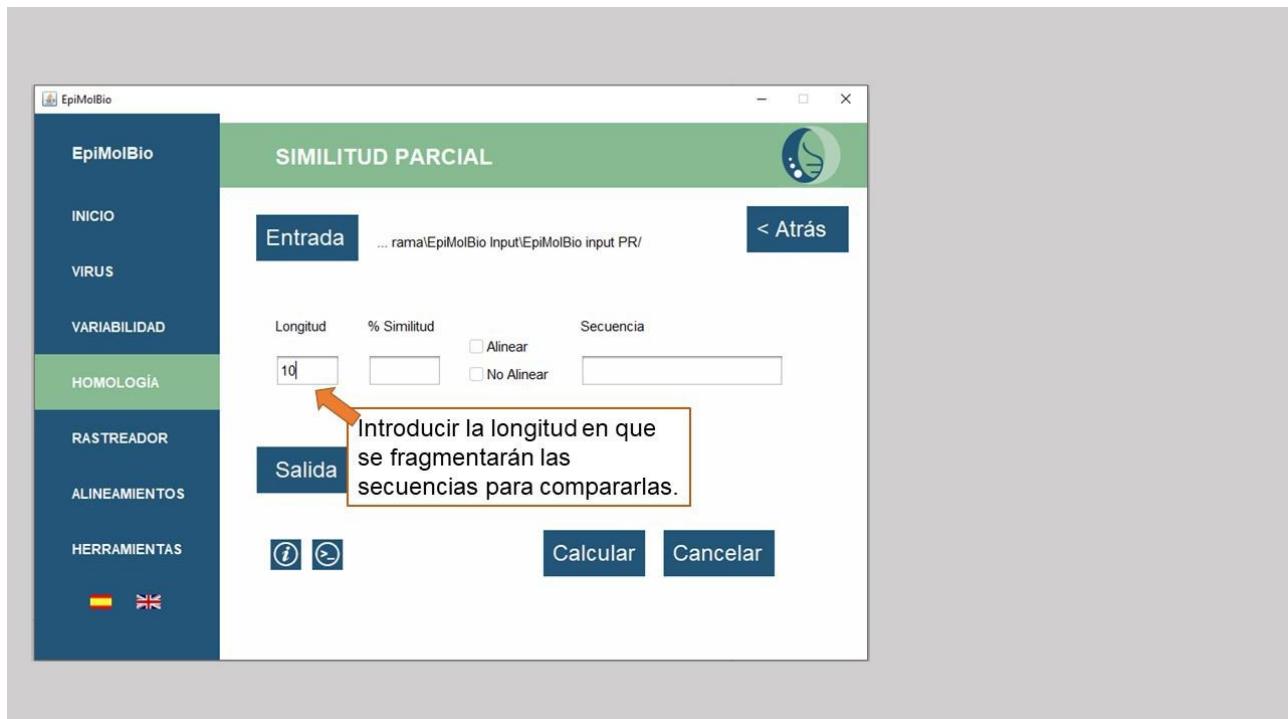
2)



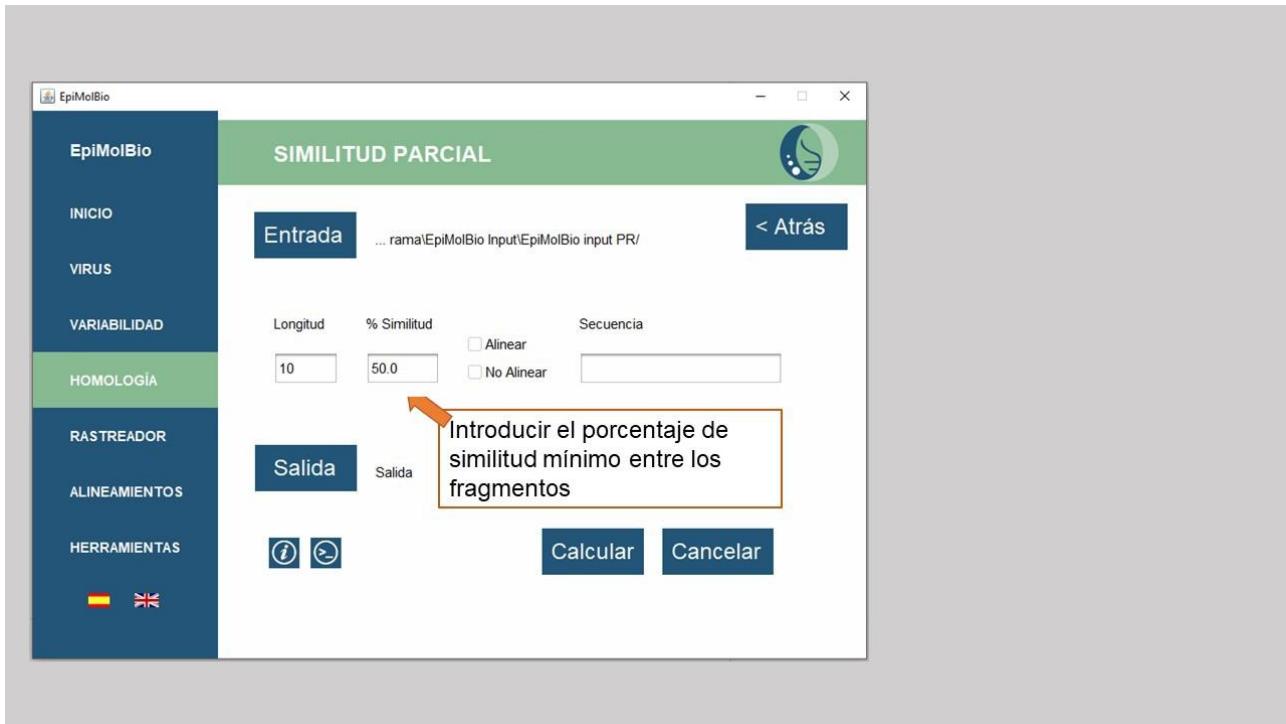
3)



4)



5)



6)



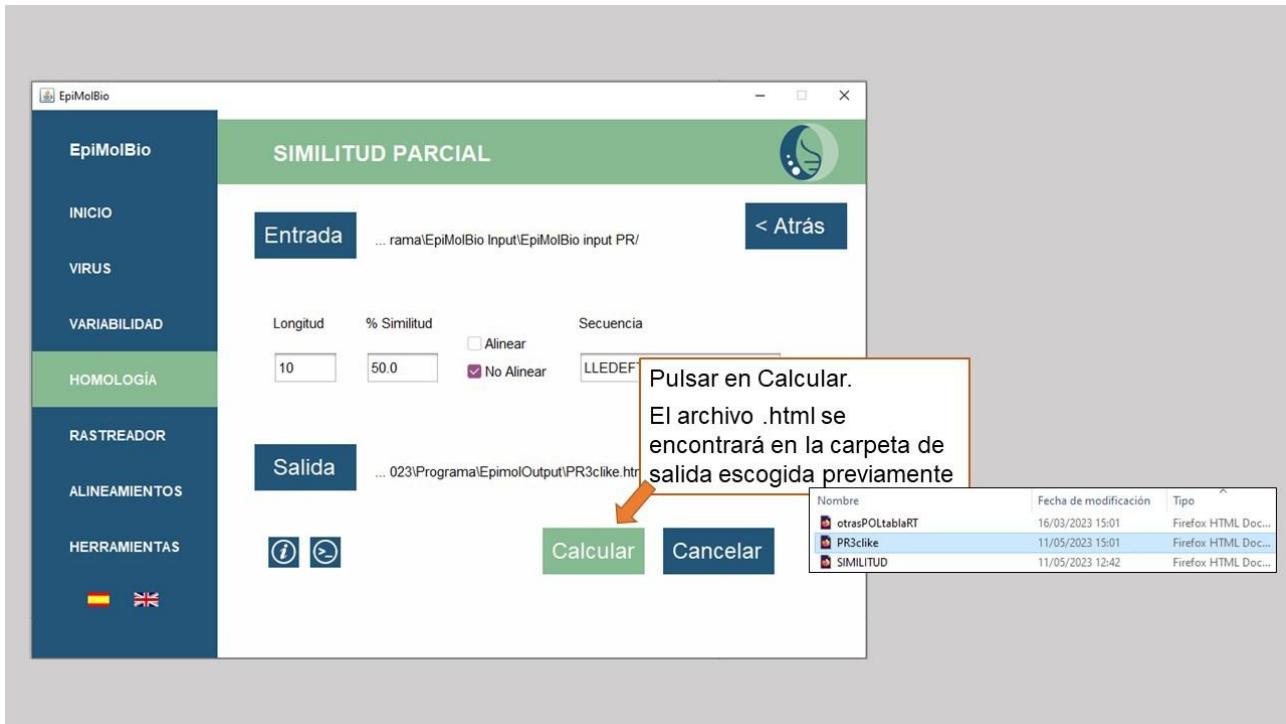
7)



8)



9)



III.3.BÚSQUEDA DE SECUENCIAS CONSERVADAS

Esta función permite obtener fragmentos de secuencias conservadas a partir de un conjunto de secuencias de entrada pudiendo buscar en una región concreta, escoger la longitud del fragmento, y el porcentaje de conservación. Por ejemplo, búsqueda de péptidos conservados para el posterior diseño de aptámeros de uso diagnóstico o terapéutico.

El formato de **salida** es una tabla .html donde se muestra, en la parte superior, el título del análisis. En la columna “Archivo”, aparece el nombre del archivo analizado, seguido de la columna “Longitud” que muestra la longitud del fragmento. En la columna “Región”, aparece la región de la secuencia de entrada donde se ha encontrado el fragmento. En la columna “Fragmento”, se muestra el resultado con la secuencia obtenida seguido de “Frecuencia”, donde aparece el porcentaje de conservación coloreado según el código de colores descrito en Generalidades que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

Ejemplo de formato de salida del análisis Homología Búsqueda de Secuencias Conservadas:

| Homología Búsqueda de Secuencias Conservadas Longitud 10 - 10 95.0% | | | | |
|---|----------|---------|------------|------------|
| Archivo | Longitud | Región | Fragmento | Frecuencia |
| PR_71_BF1.fasta | 10 | 22 - 31 | ALLDTGADDT | 100.000% |
| PR_71_BF1.fasta | 10 | 23 - 32 | LLDTGADDTV | 100.000% |
| PR_71_BF1.fasta | 10 | 24 - 33 | LDTGADDTVL | 100.000% |
| PR_71_BF1.fasta | 10 | 25 - 34 | DTGADDTVLE | 100.000% |
| PR_130_A1B.fasta | 10 | 1 - 10 | PQITLWQRPL | 100.000% |
| PR_130_A1B.fasta | 10 | 2 - 11 | QITLWQRPLV | 100.000% |
| PR_130_A1B.fasta | 10 | 3 - 12 | ITLWQRPLVT | 100.000% |

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias a analizar en nucleótidos o aminoácidos. Es recomendable que las secuencias de entrada estén alineadas y sin inserciones si se quiere buscar el fragmento conservado en una región concreta o si las secuencias de entrada no son diferentes entre sí.

En el campo “**Rango**” seleccionar “Todo el Rango” cuando se quiera buscar en toda la longitud de las secuencias de entrada o “Seleccionar Rango” cuando se quiera buscar en una región concreta de las secuencias de entrada.

Si se ha escogido “Seleccionar Rango” en el campo anterior, en el campo “**Rango Seleccionable**” introducir las posiciones de nucleótidos o aminoácidos que engloban la región donde se quiere buscar el fragmento conservado (ejemplo: para buscar entre el aminoácido 10 y el 30 inclusive, escribir en la primera caja: “10” y en la segunda, “30”).

En el campo “**% Conservación**” definir el porcentaje de conservación mínimo que queremos que tengan los fragmentos para que aparezcan en los resultados, introduciéndolo en cifras con un decimal sin el símbolo de “%” (ejemplo: 80.0 para una similitud del 80%).

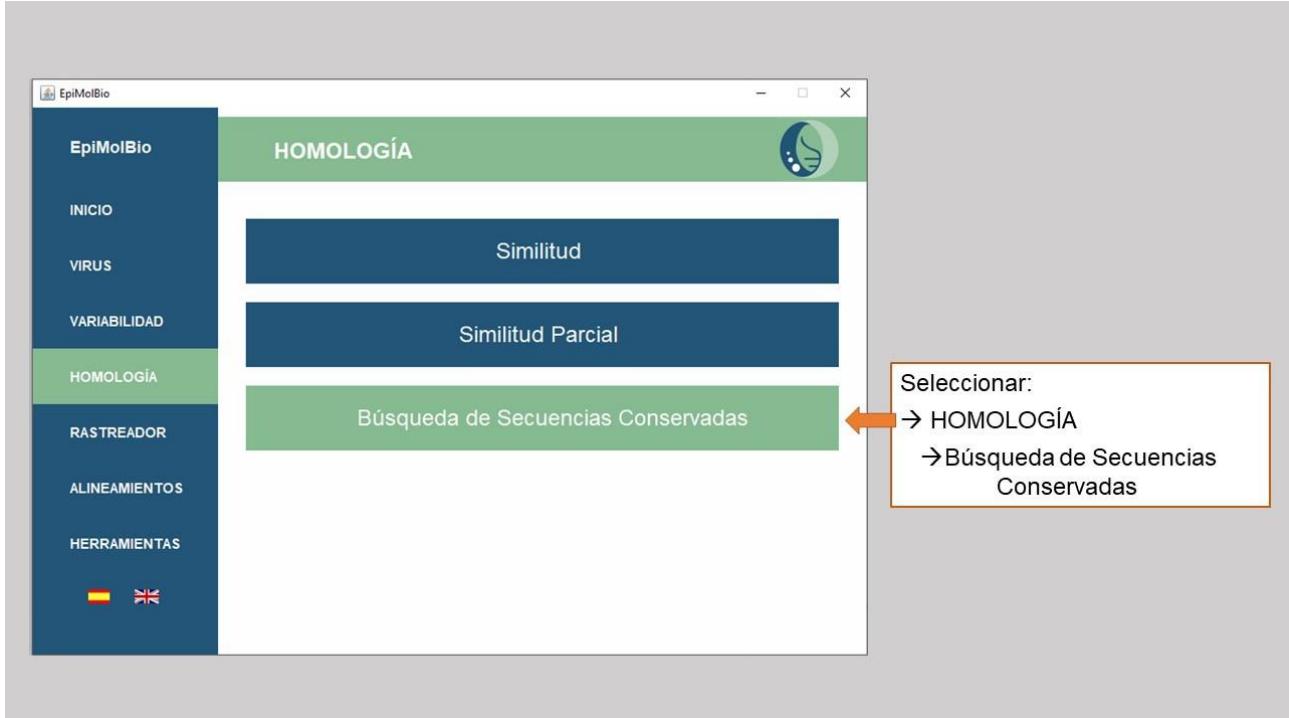
En el campo “**Longitud de Secuencias**” introducir la longitud que se quiere que tengan los fragmentos conservados resultantes. Si se desea una longitud entre 20 y 25 residuos, escribir en la primera caja “20” y en la segunda, “25”. Si se quiere que tengan 20 residuos, escribir en ambas cajas, “20”.

En el campo “**Referencia**” se puede introducir opcionalmente una secuencia de referencia para agilizar el proceso de cálculo. Ésta debe introducirse sin espacios ni saltos de línea, en nucleótidos o aminoácidos según el archivo de entrada. Si no se rellena este campo, se generará un consenso de forma automática que servirá de referencia.

En **salida** habrá que seleccionar la carpeta de salida donde se quiere que aparezcan los archivos .html y nombrar los archivos escribiendo .html al final.

Paso a paso:

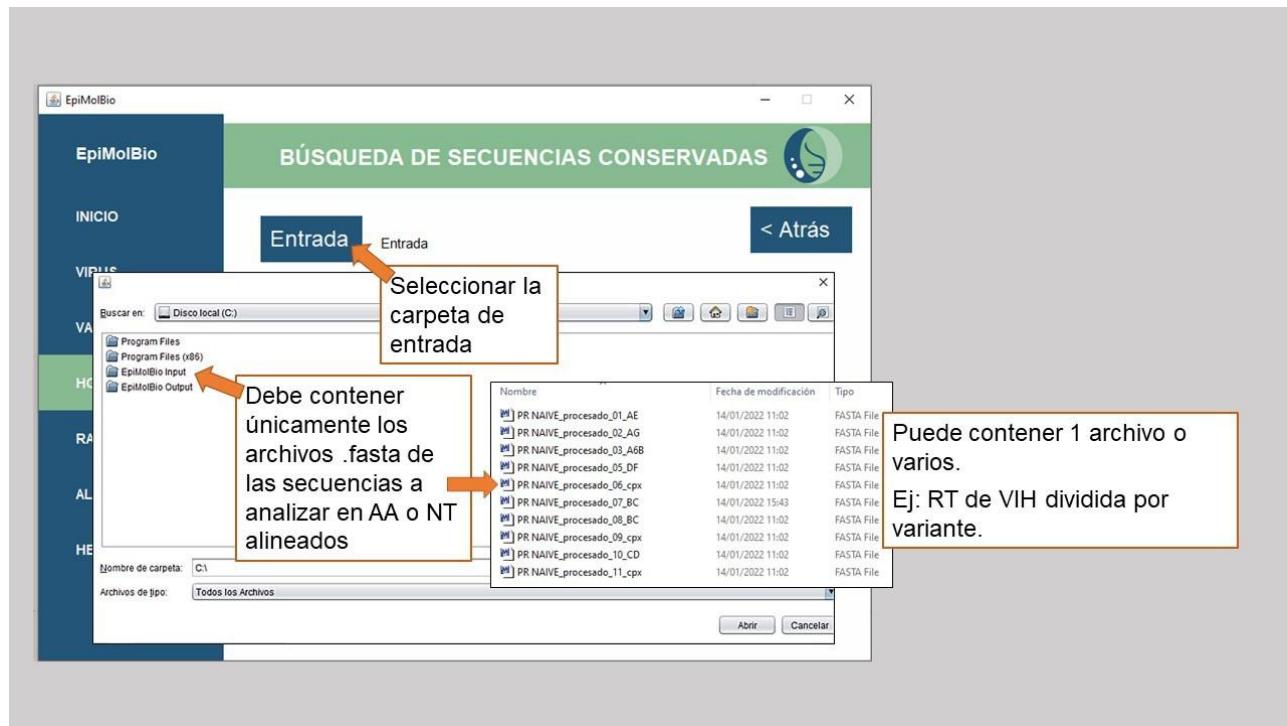
1)



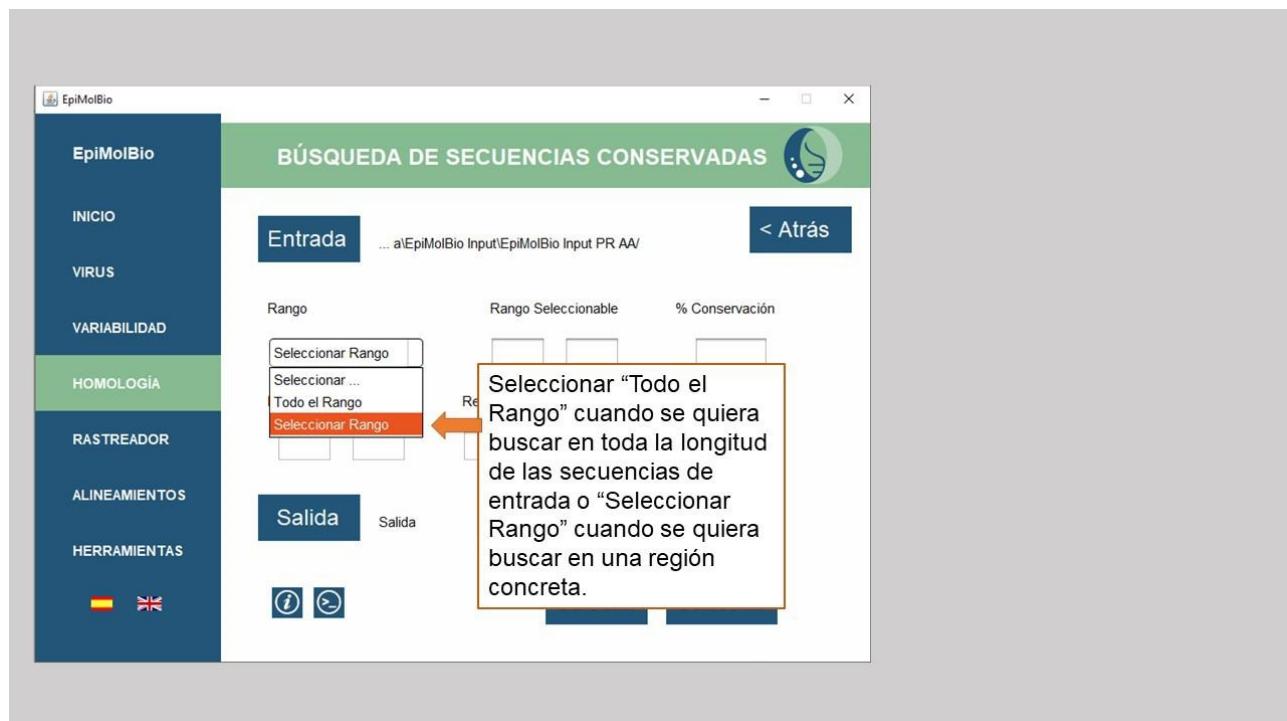
2)



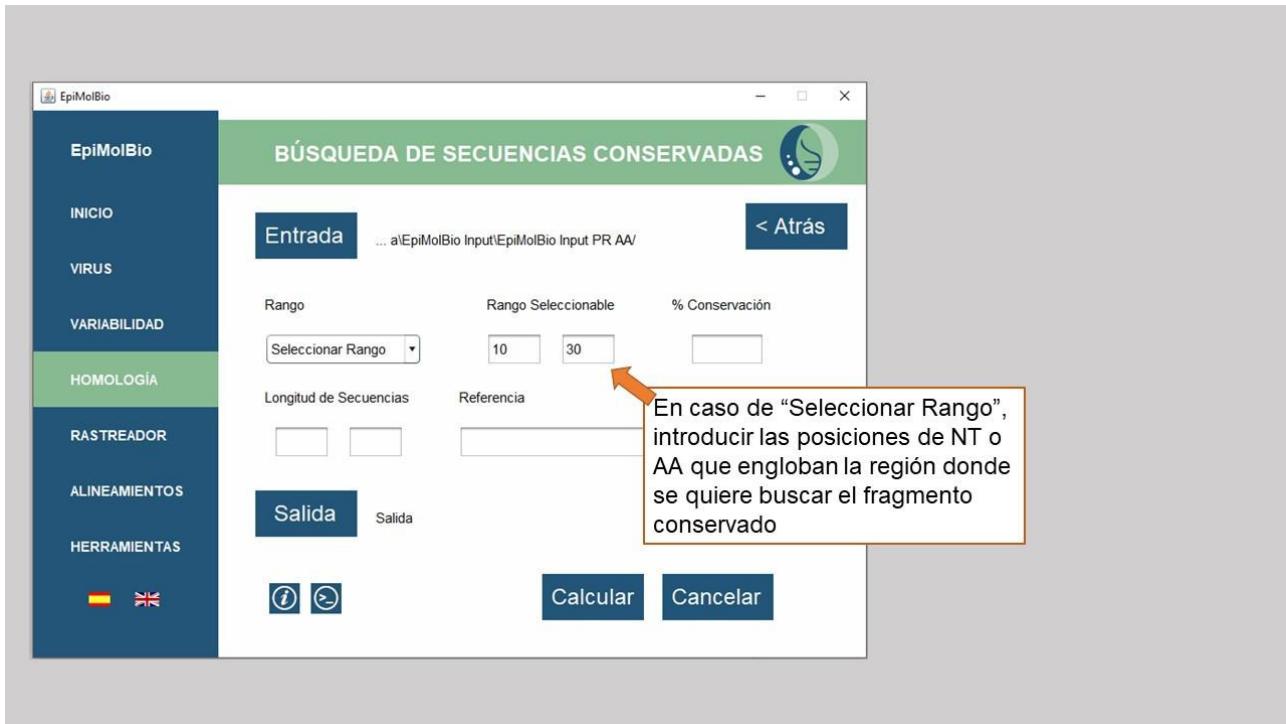
3)



4)



5)



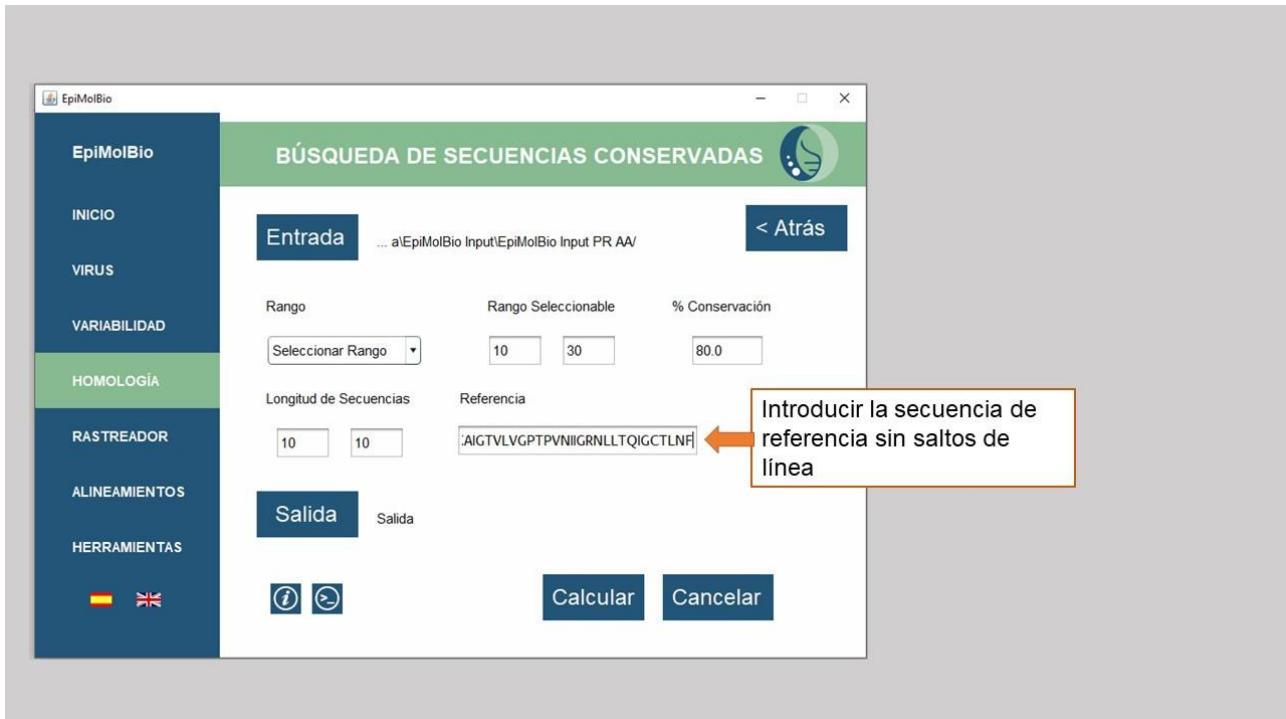
6)



7)



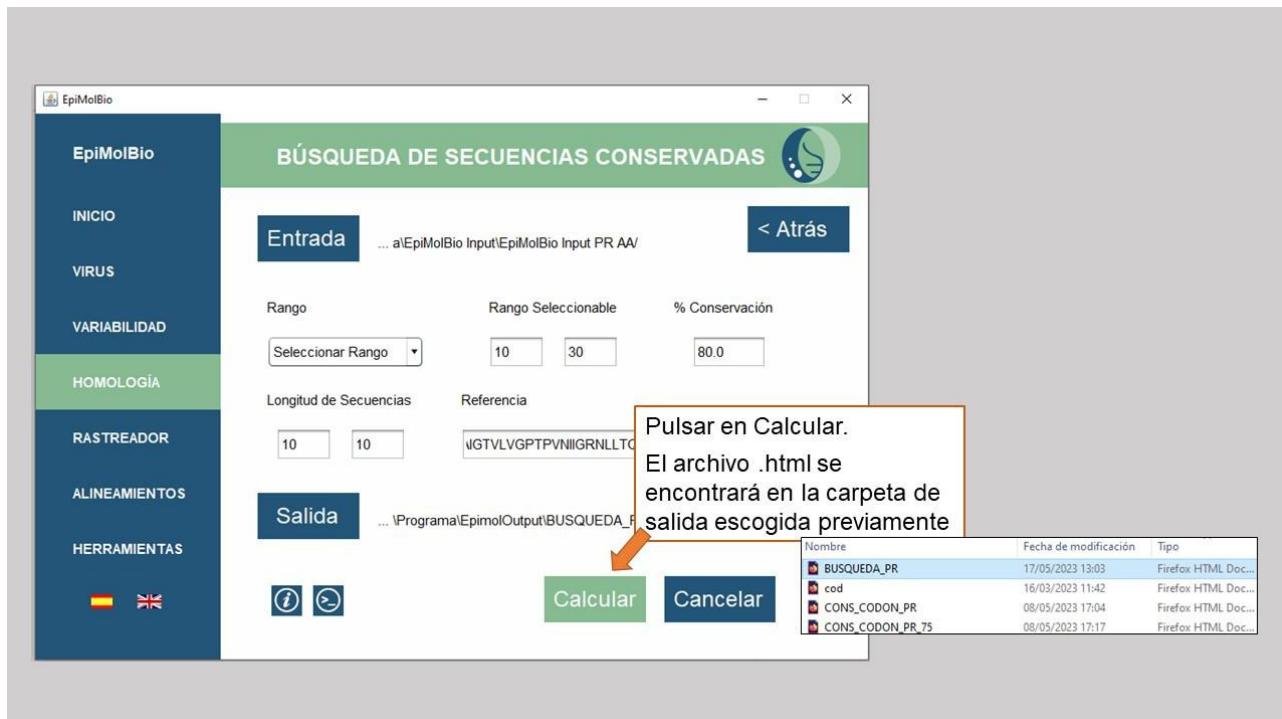
8)



9)



10)



IV.RASTREADOR

IV.1.SIMILITUD

Esta función permite buscar secuencias de interés en nucleótidos o aminoácidos dentro de un conjunto de secuencias más largas, que pueden estar incompletas o tener distintas longitudes. La búsqueda se basa en una secuencia de referencia introducida por el usuario. Por ejemplo: buscar la proteasa del VIH-2 dentro de un grupo de secuencias del genoma del VIH-2 o la secuencia de la Spike dentro del genoma completo de SARS-CoV-2. Se puede escoger el porcentaje de similitud entre lo que se quiere encontrar y la secuencia de referencia para que no se descarten secuencias válidas por presentar mutaciones. El formato de **salida** es un archivo en formato .fasta con las secuencias encontradas.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias, preferiblemente completas, donde se quiera buscar. Las secuencias pueden estar en nucleótidos (NT) o aminoácidos (AA), no estar alineadas y tener distintas longitudes.

Seleccionar si se quiere **traducir o no** la secuencia resultante. Si la entrada está en NT y se quiere la salida en NT, seleccionar “No Traducir” (la referencia debe estar en NT). Si la entrada está en NT y la salida se quiere en AA, seleccionar “Traducir” (la referencia debe estar en AA). Si la entrada está en AA, seleccionar “No Traducir”, la salida estará en AA (la referencia también debe estar en AA).

Seleccionar “**Rango Completo**” si se quiere buscar en toda la longitud de las secuencias de entrada, o bien “**Seleccionar Rango**” para buscar en una región concreta. En este último caso, las secuencias de entrada deben estar alineadas.

En caso de haber escogido “Seleccionar Rango” en el campo anterior, introducir las posiciones de nucleótidos o aminoácidos que engloban la región donde se quiere buscar el fragmento (ej.: para buscar entre el AA 10 y el 30 inclusive, escribir en la primera caja: “10” y en la segunda, “30”) en el campo “**Rango**”.

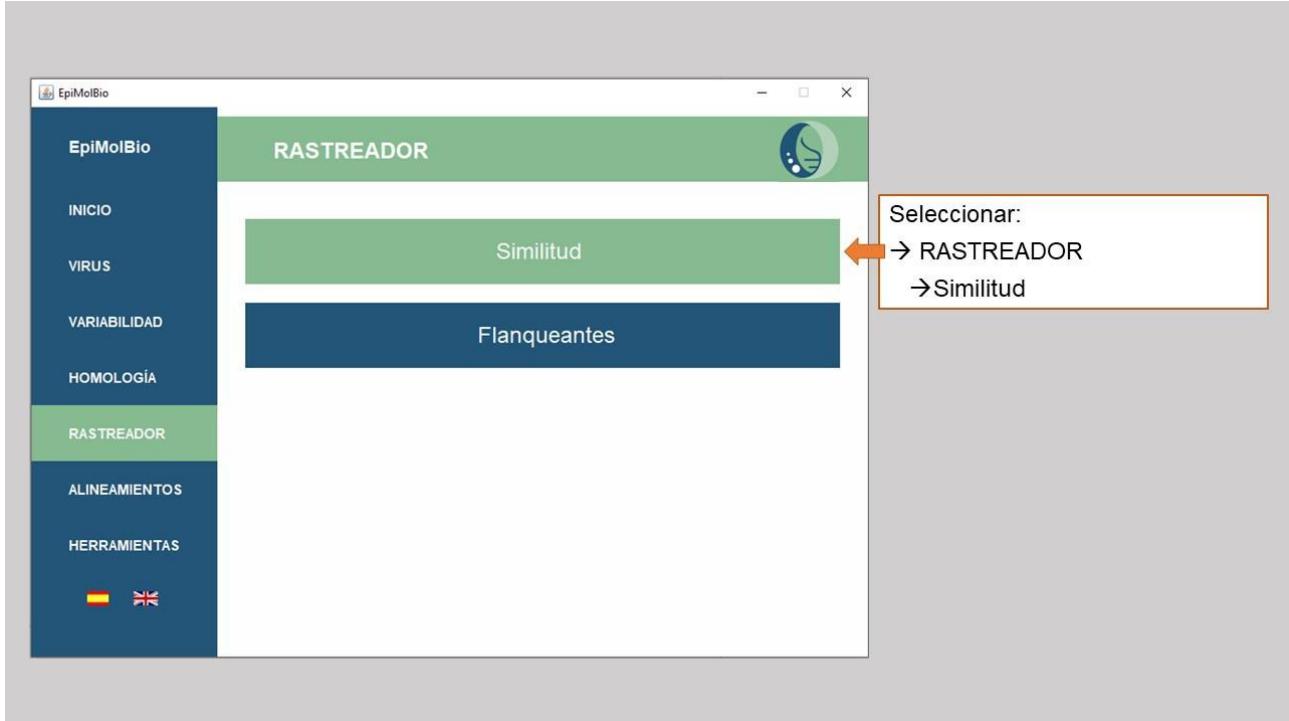
En el campo “**% Similitud**”, definir el porcentaje de similitud mínimo que queremos que tengan las secuencias de salida con respecto a la secuencia de referencia, introduciéndolo en cifras con un decimal sin el símbolo de “%” (ej.: 90.0 para una similitud del 90%).

En el campo “**Referencia**”, introducir la secuencia de referencia de la secuencia que buscamos sin saltos de línea ni espacios (por ejemplo, la secuencia de referencia del Spike cuando se busca dentro del genoma completo del SARS-CoV-2).

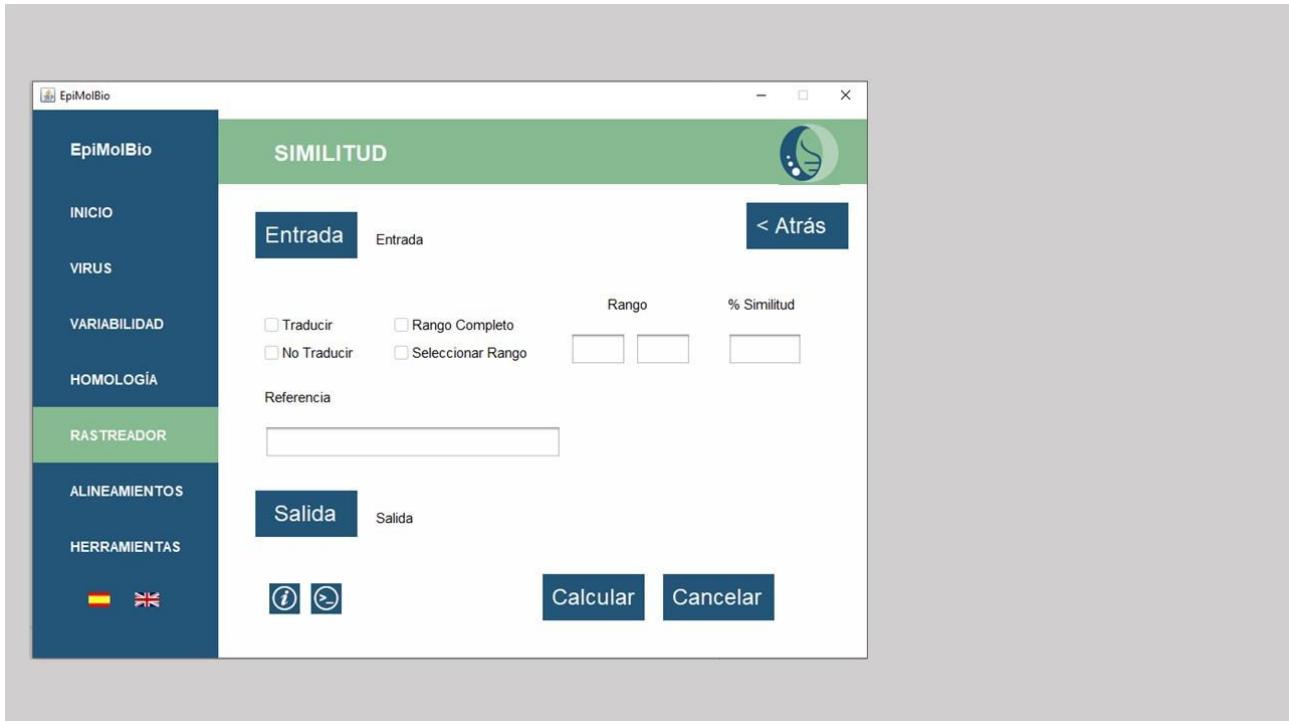
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .fasta con las secuencias encontradas. Los archivos se nombran de forma automática de la siguiente forma: Rastreado_Similitud_Nombre del archivo de entrada.fasta

Paso a paso:

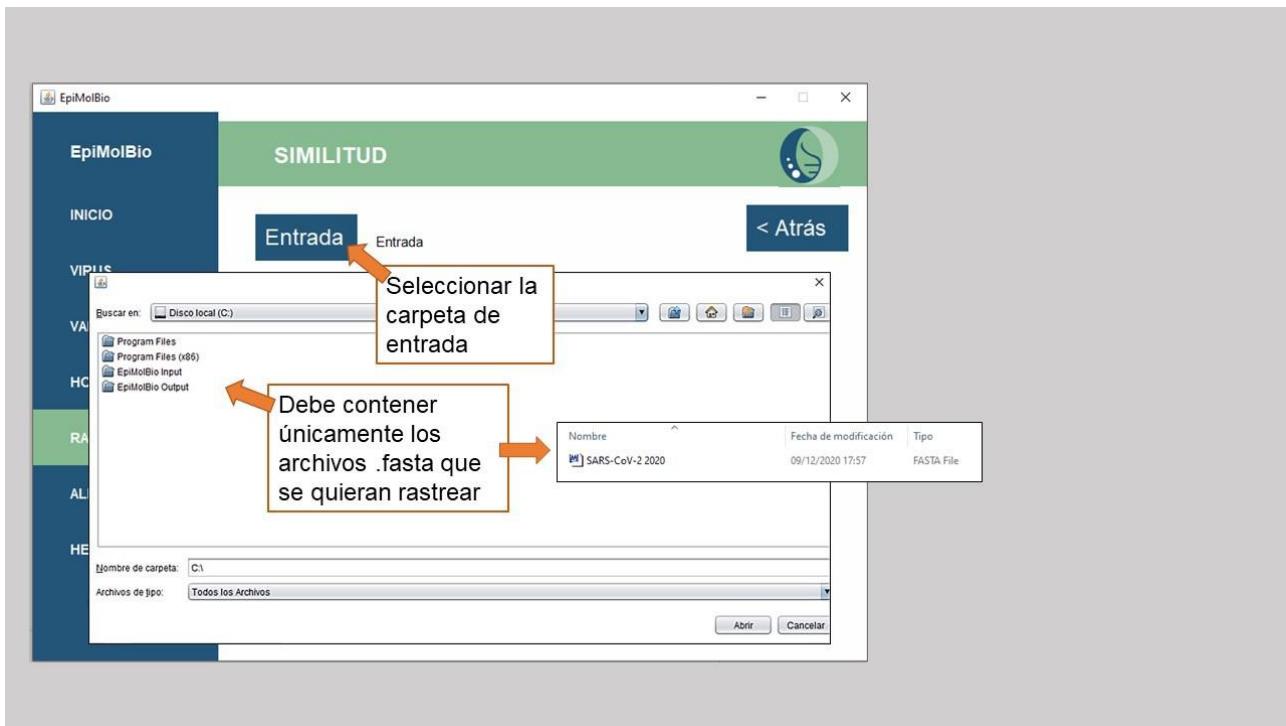
1)



2)



3)



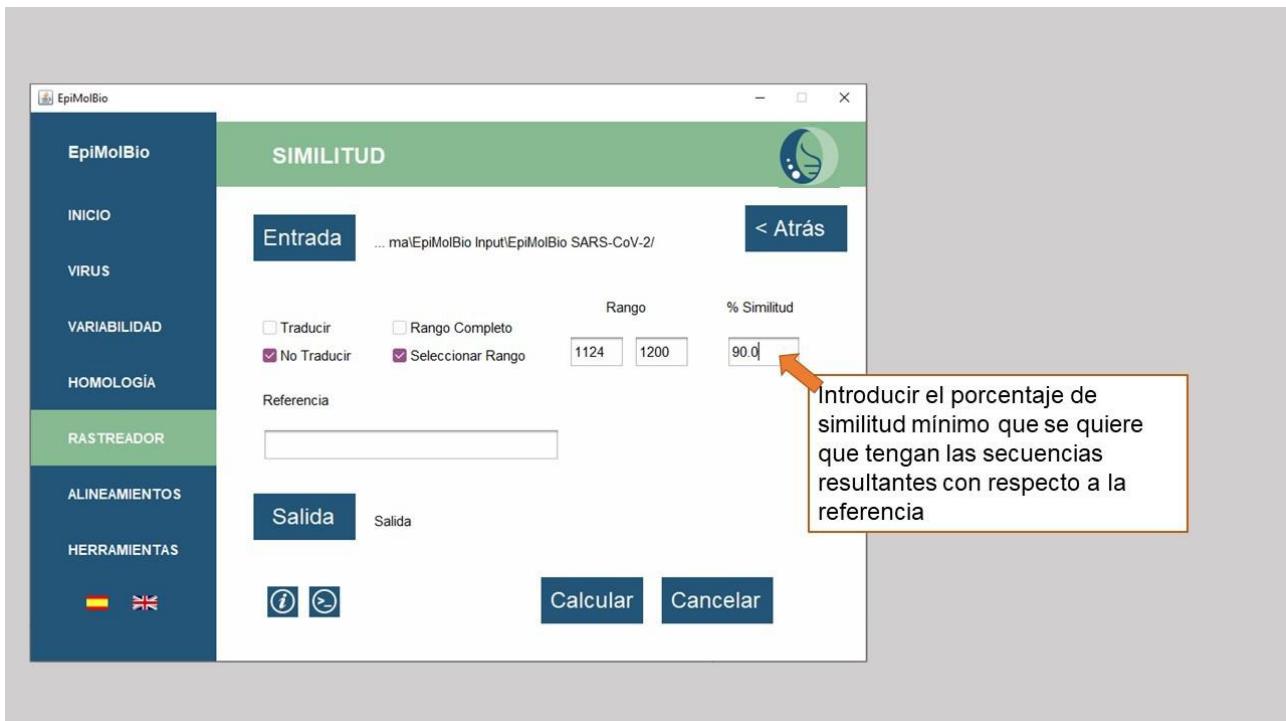
4)



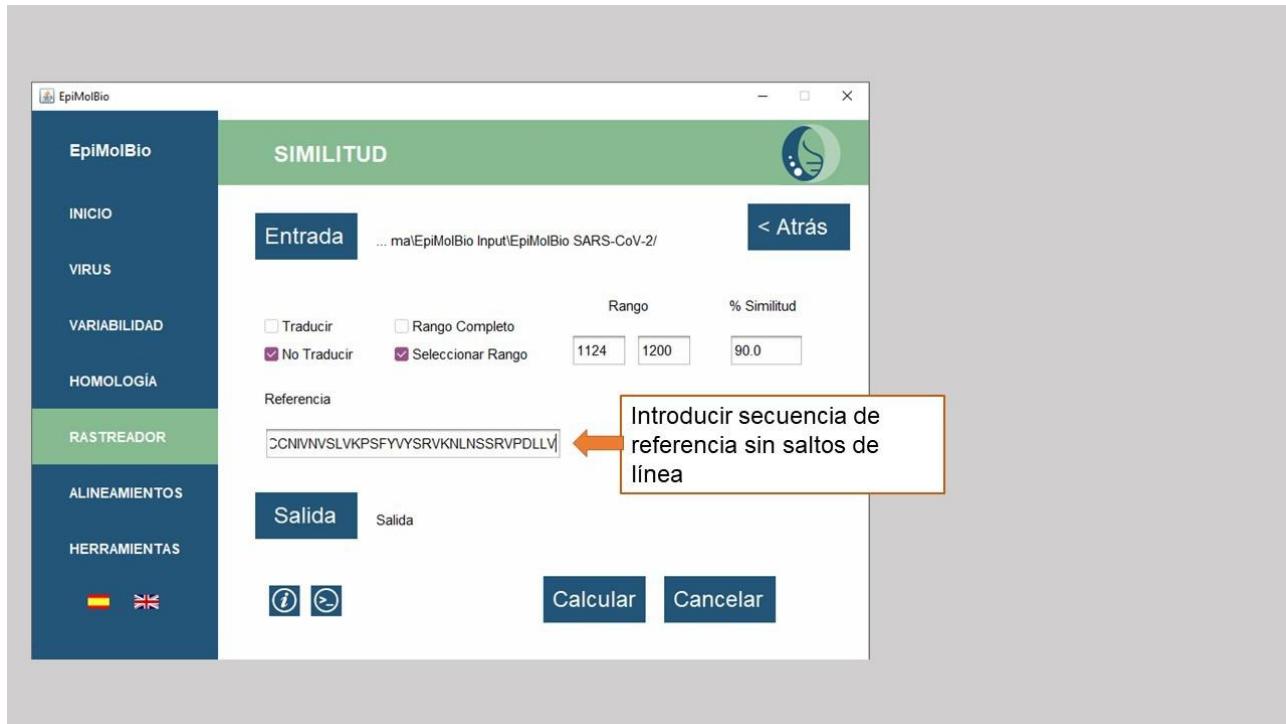
5)



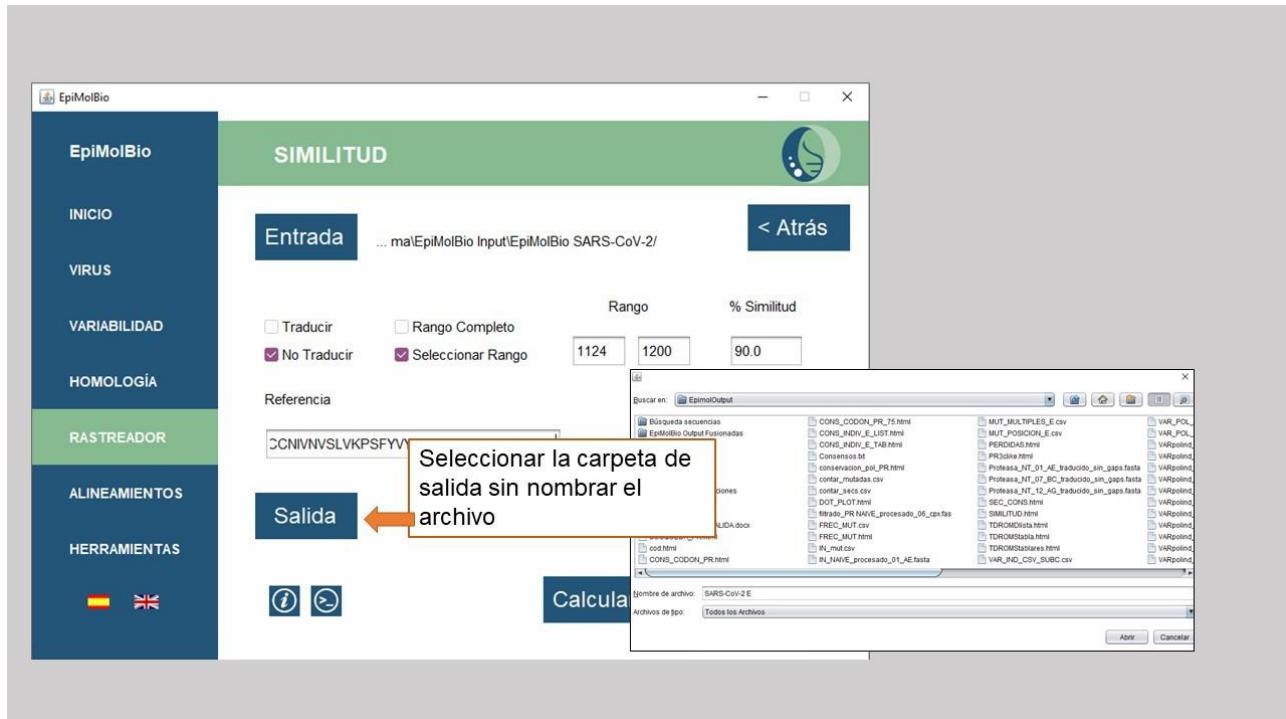
6)



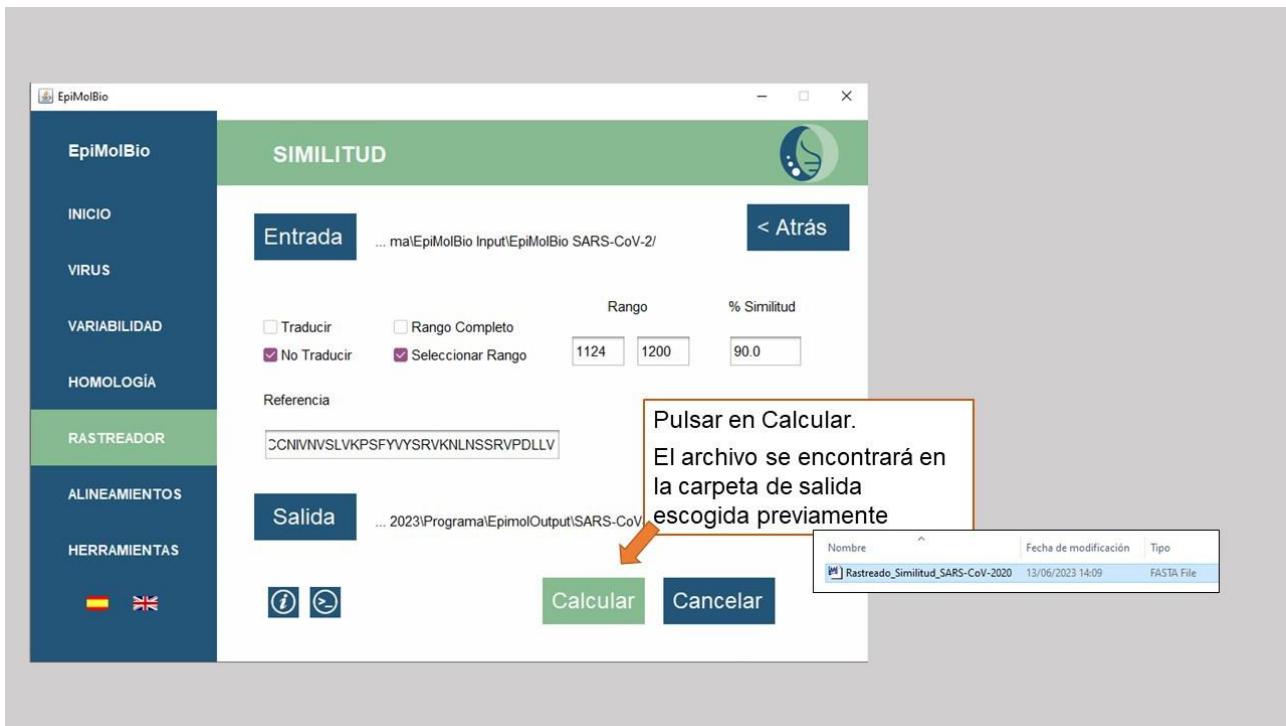
7)



8)



9)



IV.2.FLANQUEANTES

Esta función **permite buscar proteínas dentro de un conjunto de secuencias genómicas completas utilizando secuencias flanqueantes (secuencias anteriores y posteriores de unos 15 residuos) a la secuencia proteica buscada.** Estas secuencias flanqueantes deben estar en aminoácidos y en el mismo marco de lectura que la secuencia que se quiere buscar. Para encontrar la proteína de interés, ésta debe estar completa dentro del archivo de entrada. Las secuencias incompletas no se podrán localizar. El formato de salida es un archivo en formato .fasta con las secuencias encontradas.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos **.fasta** con las secuencias completas en nucleótidos donde se quiera buscar. Las secuencias de entrada pueden no estar alineadas o parcialmente incompletas, pero siempre deben tener longitudes similares ya que la búsqueda se hará por rangos de localización de residuos. Si las secuencias no son similares, los rangos no corresponderán a las regiones donde se encuentre la proteína de interés, y el programa realizará la búsqueda incorrectamente.

En el campo “**Tipo de secuencia**”, seleccionar aminoácidos o nucleótidos según si se quiere que el archivo de salida esté o no traducido.

En el campo “**Rango**” introducir en cifras el rango de la secuencia genómica donde se quiere buscar la proteína de interés. Por ejemplo, para buscar la proteína de la envuelta en el genoma del SARS-CoV-2, introducir 26050 en la primera caja de rango y 26650 en la segunda (las cifras deben introducirse sin puntos). El programa buscará entre los **nucleótidos** 26.050 y 26.650 de los genomas del archivo de entrada.

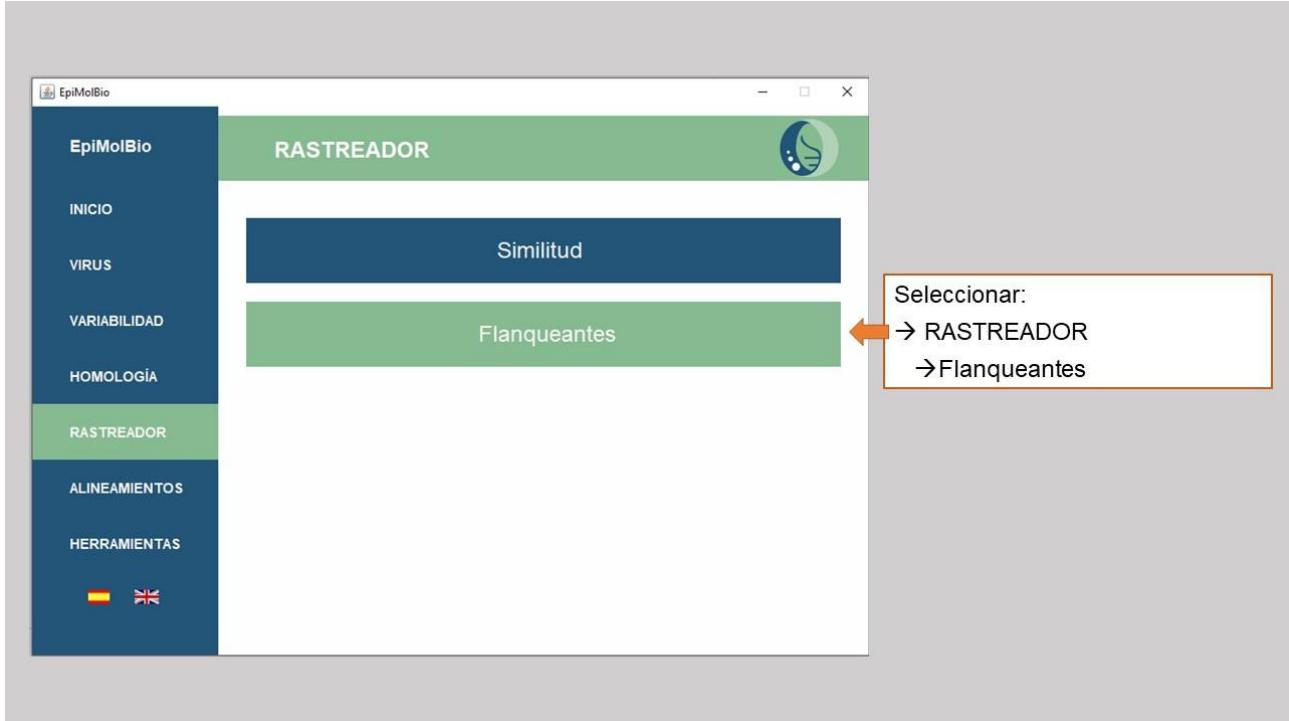
En el campo “**Tamaño**” introducir en cifras la **longitud** de proteína que se está buscando en aminoácidos. En este ejemplo sería 75.

En el campo “**Secuencias Flanqueantes**” introducir los **15 aminoácidos** anteriores y posteriores a la proteína de interés que se está buscando. Por ejemplo introducir TTSVPL*AQADEYEL en la primera caja y TN*ILY*FFCLEL*F en la segunda caja de “Secuencias flanqueantes”.

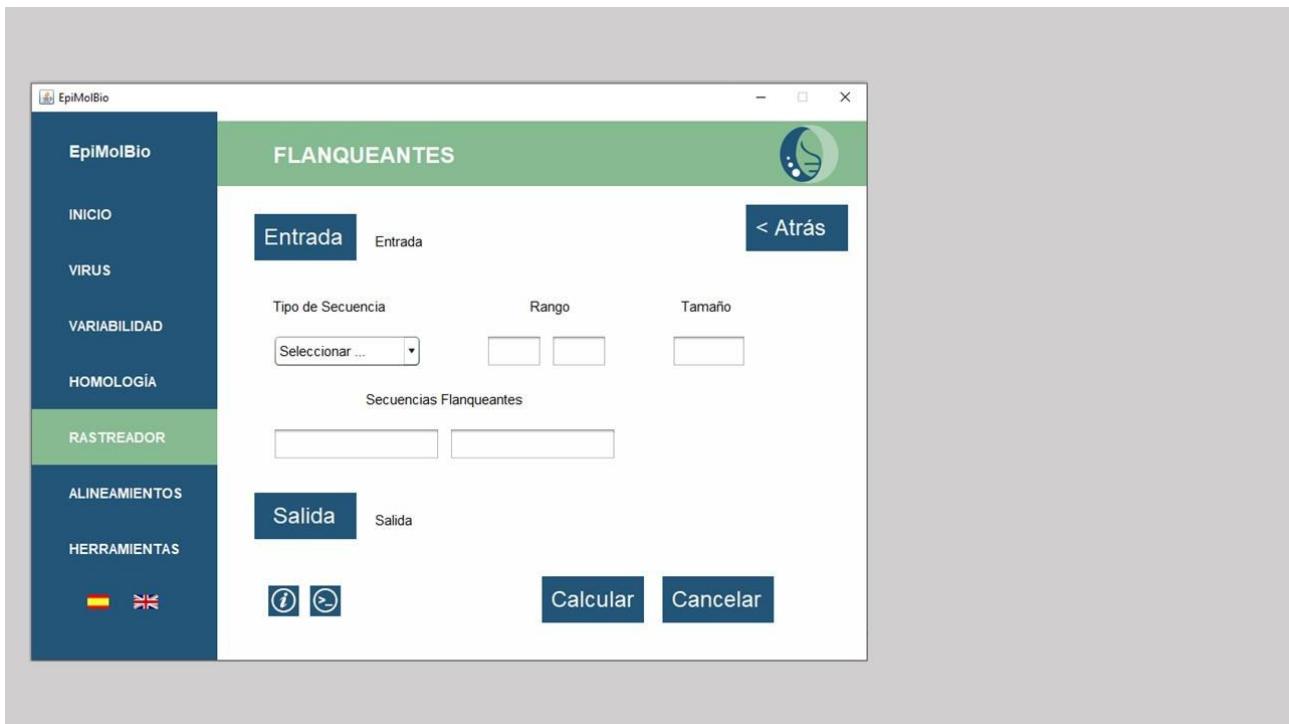
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos **.fasta**. Los archivos se nombran de forma automática de la siguiente forma: “Rastreado_Flanqueantes_Nombre del archivo de entrada.fasta”.

Paso a paso:

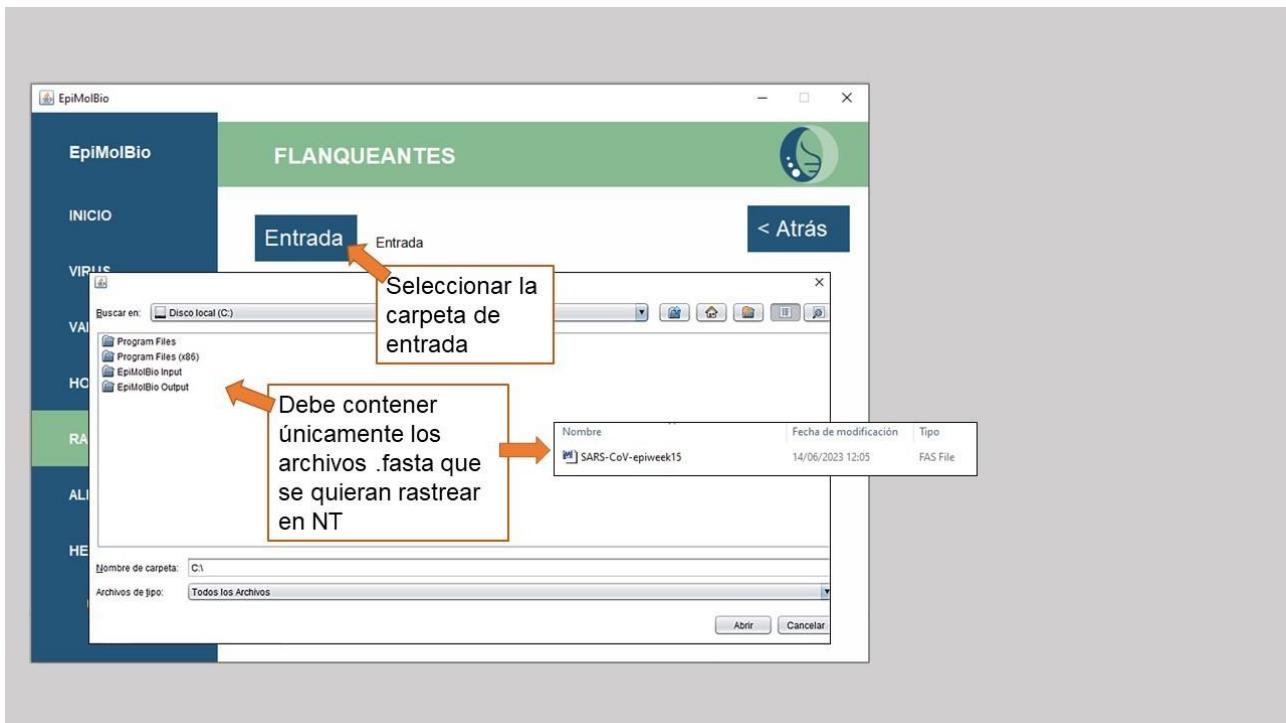
1)



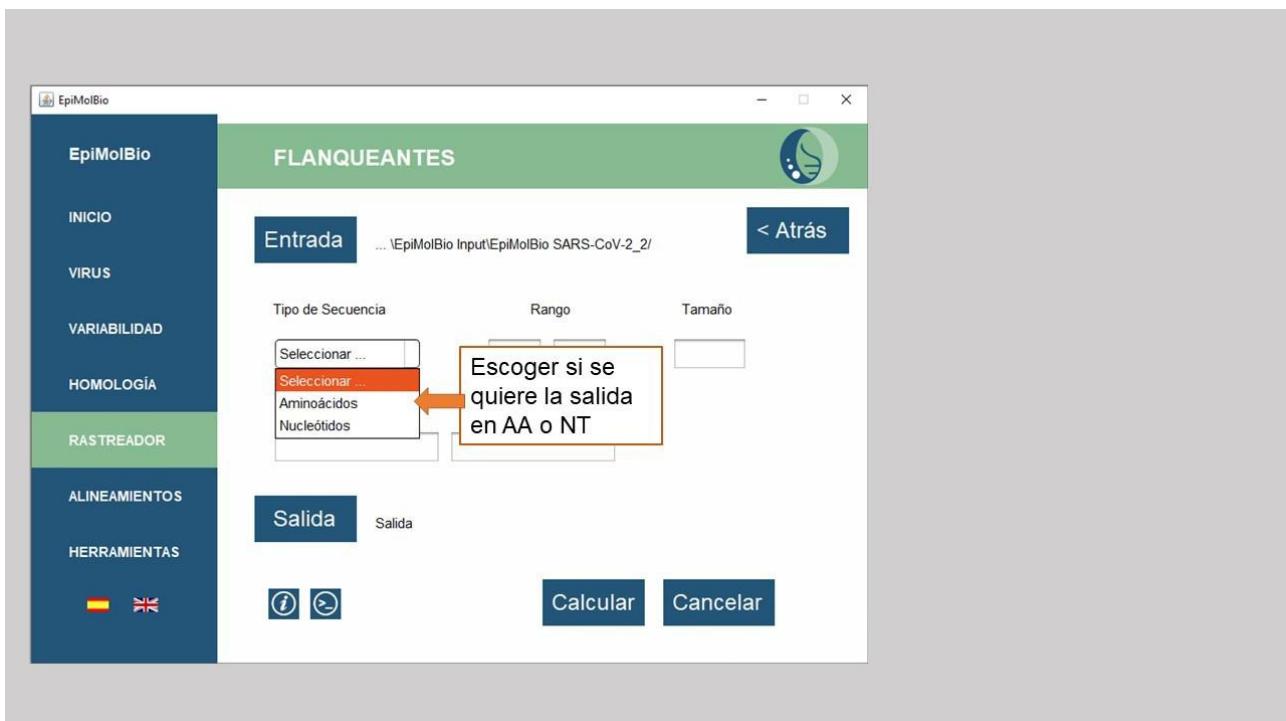
2)



3)



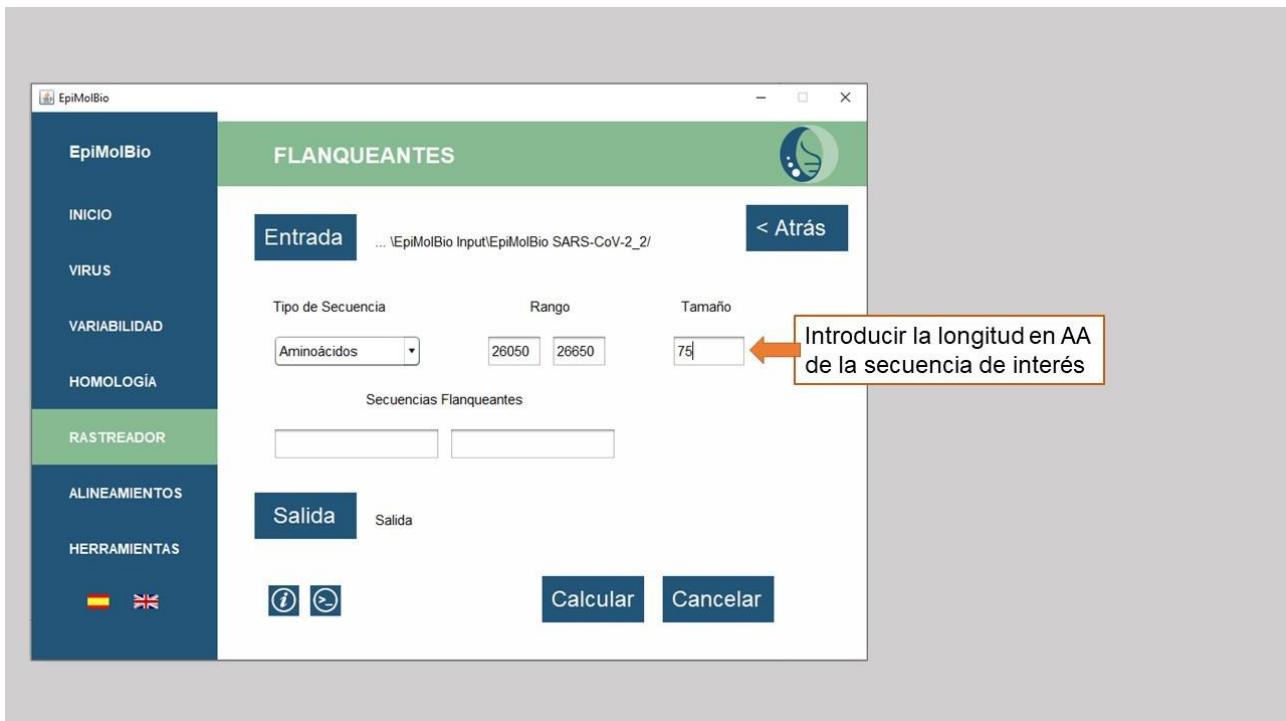
4)



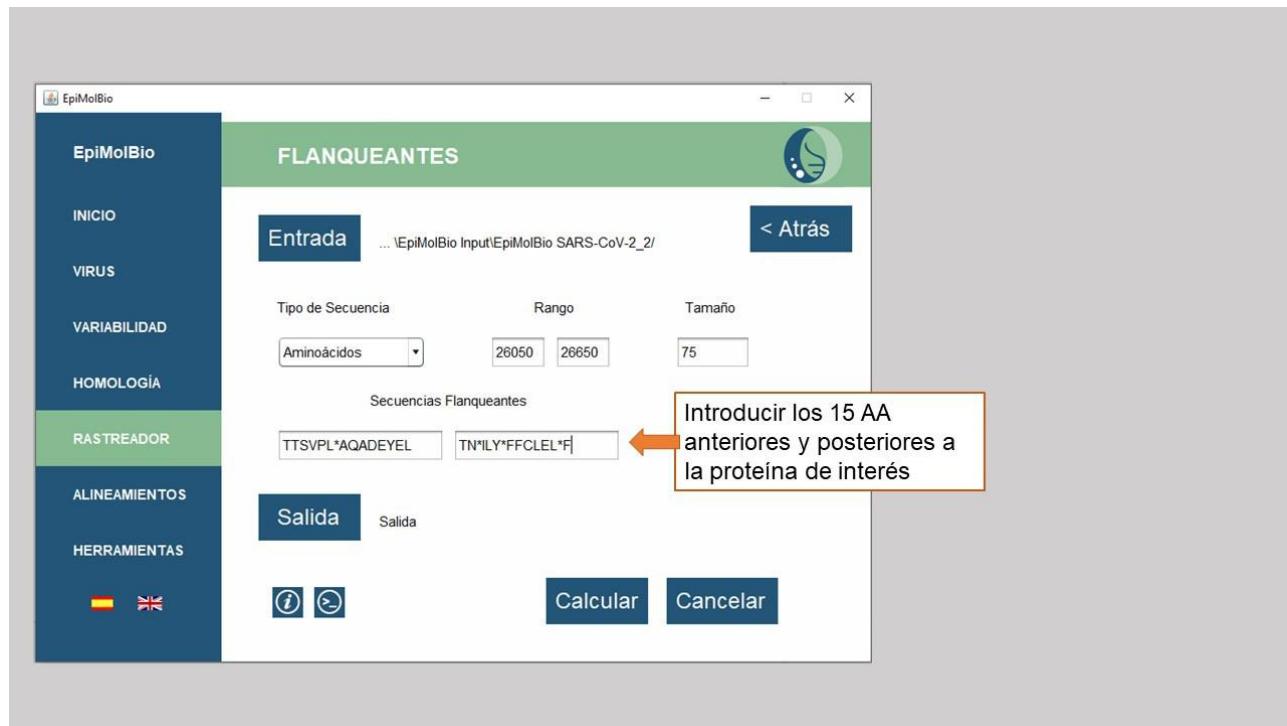
5)



6)



7)



8)



9)



V. ALINEAMIENTOS

V.1. ALINEAMIENTOS MÚLTIPLES

Esta función permite alinear secuencias de aminoácidos y nucleótidos mediante el programa de dominio público **MUSCLE v3.8.31** de Robert C. Edgar: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97. El alineamiento se lleva a cabo con respecto a una secuencia de referencia introducida por el usuario, a excepción de que se marque “Conservar Inserciones”, en este caso se hará un alineamiento múltiple entre las secuencias de entrada.

MUSCLE tiene la limitación de no poder alinear más de unos pocos miles de secuencias no demasiado largas, pero con esta función, EpiMolBio puede alinear las secuencias que se requieran, acelerando el proceso y permitiendo el alineamiento simultáneo de miles de secuencias. Para ello, se puede escoger entre alinear las secuencias en paquetes o alinearlas de una en una, en ambos casos con respecto a una referencia.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos **.fasta** en aminoácidos o nucleótidos con las secuencias que se quiera alinear.

Seleccionar el campo “**Conservar Inserciones**”, si se desea conservarlas. De lo contrario el programa las elimina de forma automática. Si se marca esta opción, no habrá que introducir nada más excepto la salida.

En el campo “**Secuencias por Archivo**” se debe introducir un número para que el programa divida el archivo de entrada en archivos con menos secuencias y MUSCLE pueda alinearlos. Se recomienda introducir en este campo el **valor de 1** en caso de querer alinear **secuencias incompletas** con respecto a la referencia o cuando el **número de secuencias sea elevado y su longitud relativamente grande**.

En caso de que las secuencias tengan muchas **mutaciones** (inserciones, delecciones y cambios de residuos) con respecto a la referencia y el **número** de secuencias y la **longitud** de estas **no sea demasiado grande**, se recomienda introducir en “**Secuencias por Archivo**” el **número total de secuencias** que tenga el archivo de entrada.

Si las secuencias tienen muchas **mutaciones** (inserciones, delecciones y cambios de residuos) y el **archivo a alinear es muy pesado** debido a que las secuencias sean muy largas o haya un alto número de secuencias (pocas secuencias pero muy largas o muchas secuencias cortas) es preferible realizar un **alineamiento múltiple** dividiendo el archivo original en paquetes con menor número de secuencias. Para ello hay que introducir un valor entre **100-500** en “**Secuencias por Archivo**”. De esta manera, se divide el archivo de entrada en paquetes con el número de secuencias introducido en el programa. Los paquetes se alinearán individualmente con respecto a la secuencia de referencia, se eliminarán los gaps y se generará un archivo unificado con las secuencias alineadas.

En el campo “**Referencia**”, introducir la secuencia de referencia sin saltos de línea en aminoácidos o nucleótidos, en NT o AA según el formato del archivo de entrada.

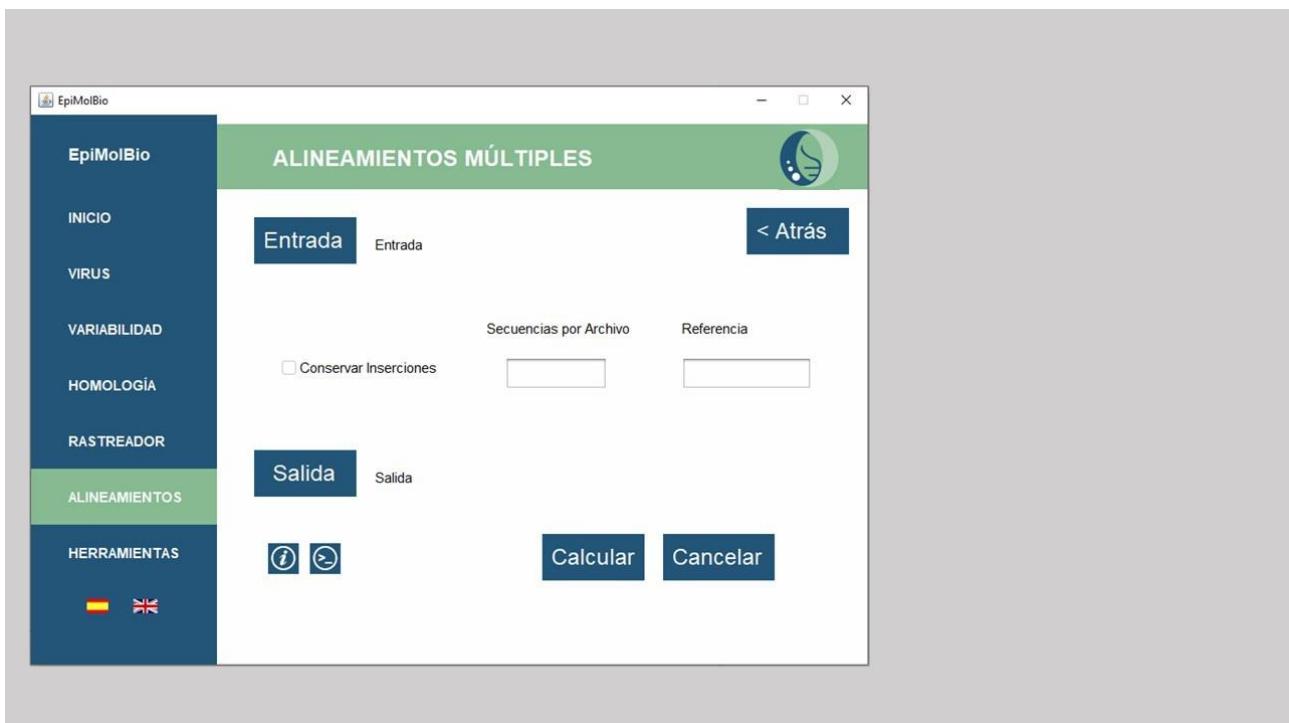
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos **.fasta** de las secuencias alineadas. Los archivos se nombran de forma automática de la siguiente forma: Alineado_Nombre del archivo de entrada.fasta.

Paso a paso:

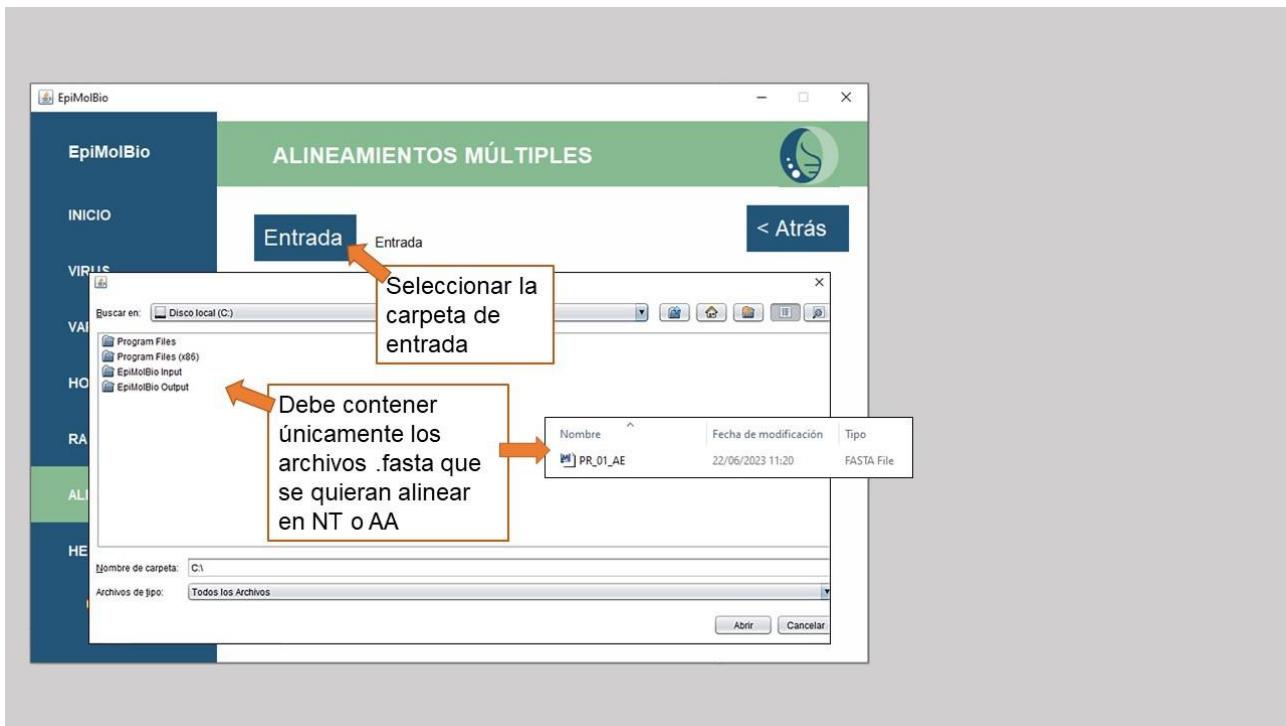
1)



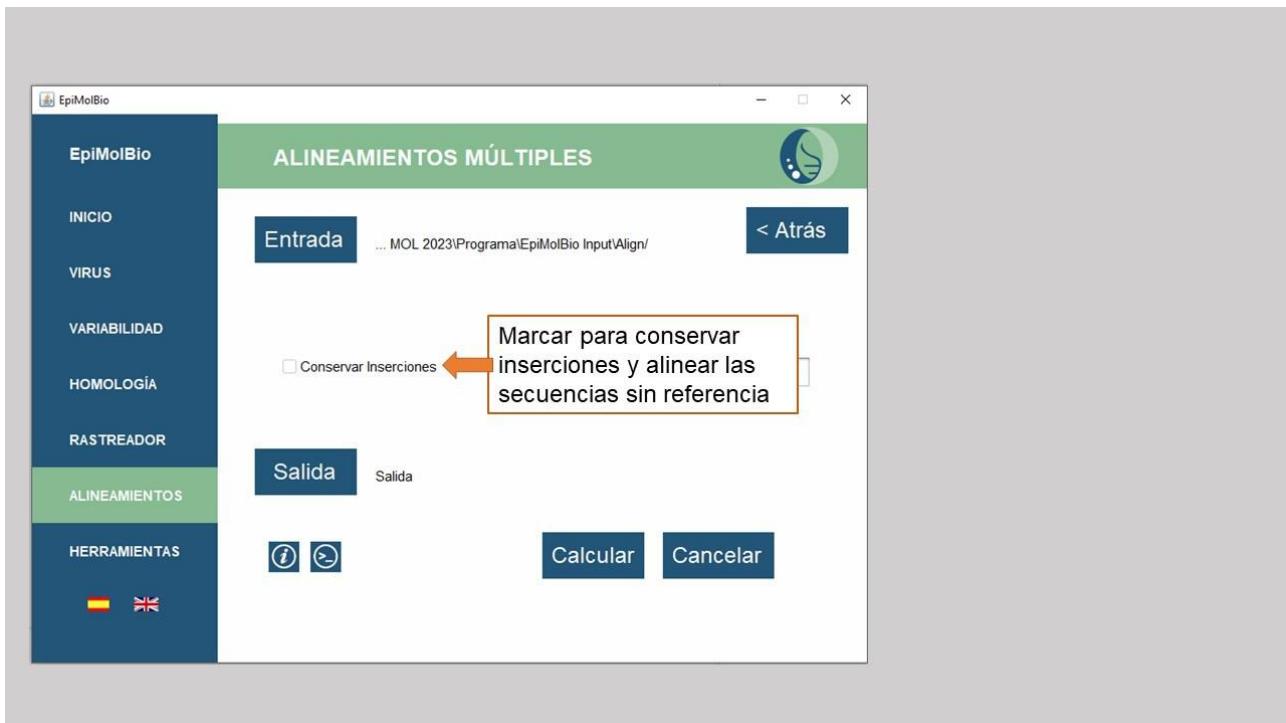
2)



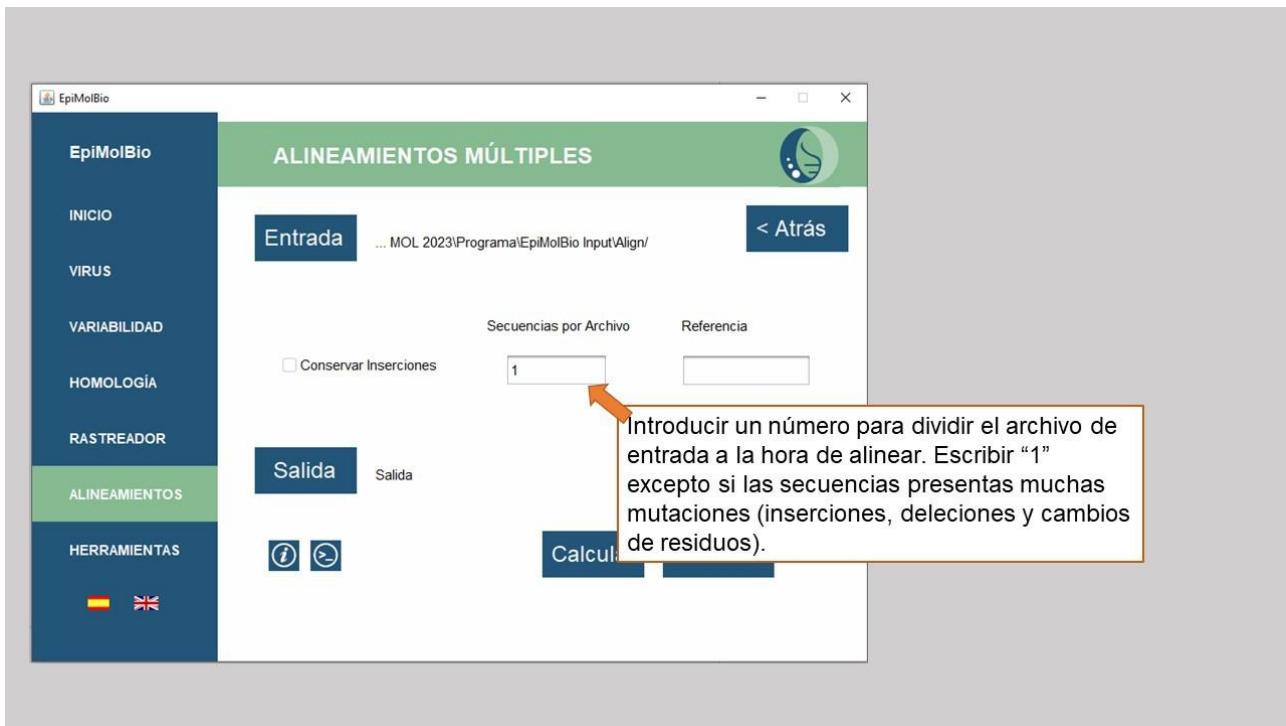
3)



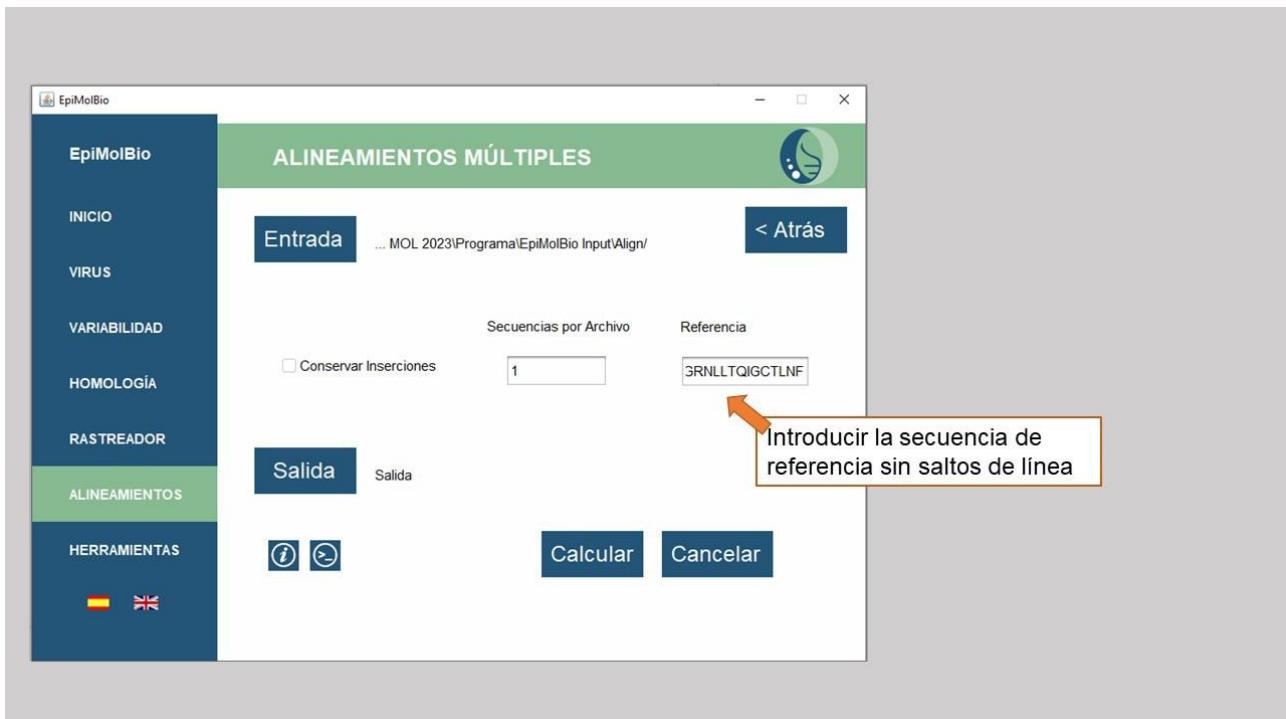
4)



5)



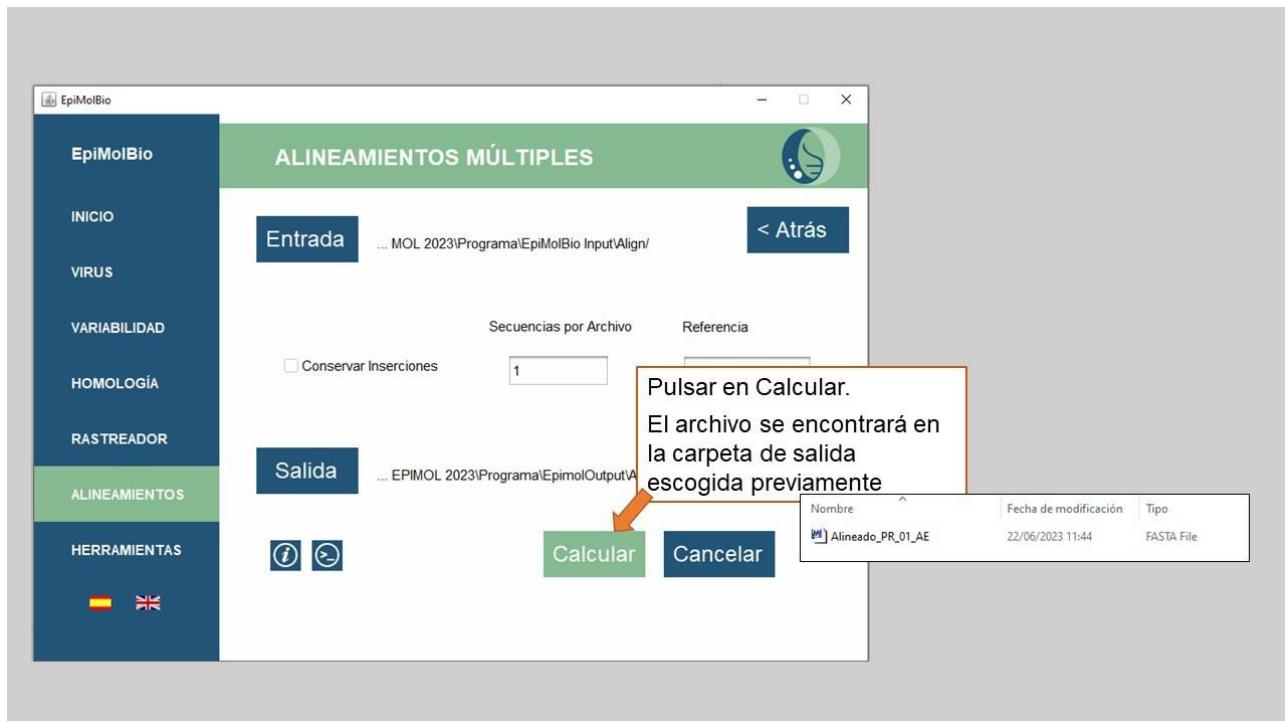
6)



7)



8)



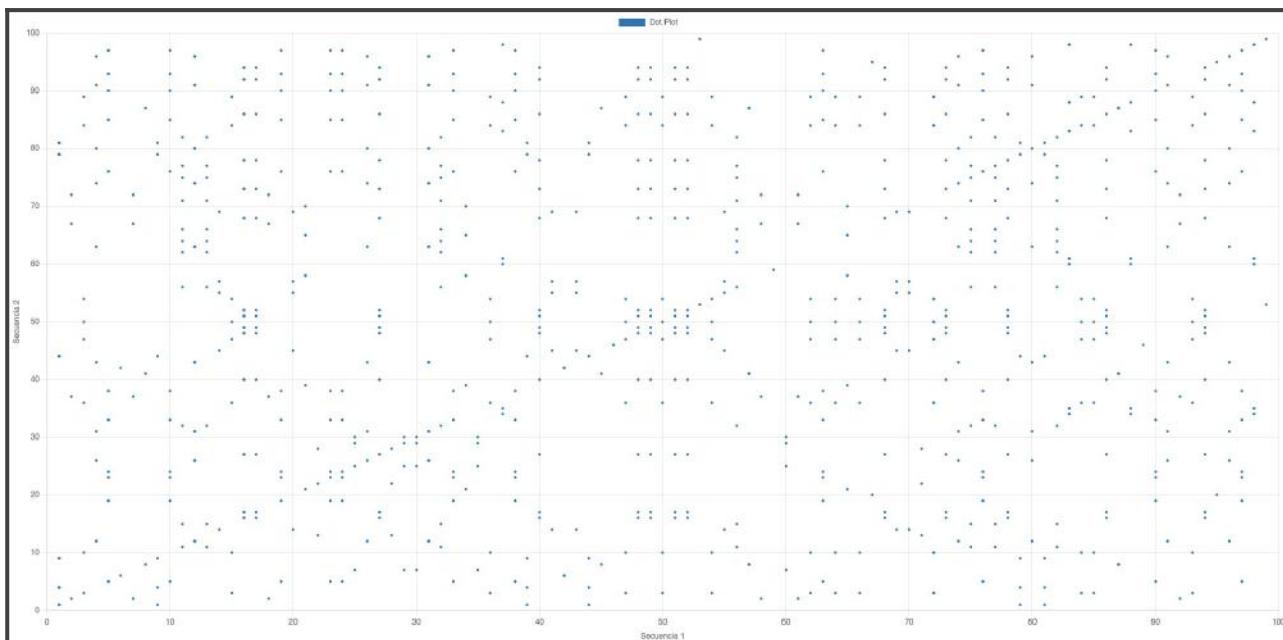
V.2.DOT PLOT

Esta función permite obtener una representación gráfica donde se comparan secuencias de ADN o proteínas. El dot plot se construye trazando puntos en una matriz bidimensional, donde cada eje representa una secuencia. Cada punto en el dot plot se coloca en la posición (x, y) cuando los residuos correspondientes en las secuencias coinciden en esas posiciones. Si hay una coincidencia, se coloca un punto; de lo contrario, se deja vacío.

Al observar el dot plot, se pueden identificar patrones como regiones de alta similitud o repeticiones en las secuencias. Los dot plots también pueden ayudar a detectar inversiones, delecciones o inserciones en las secuencias.

El formato de **salida** es un gráfico en formato .html con las secuencias comparadas, indicando las posiciones de ambas secuencias de 10 en 10.

Ejemplo de formato de salida del análisis Dot Plot:



En el campo “**Secuencia 1**” introducir la primera secuencia a comparar en nucleótidos o aminoácidos sin saltos de línea ni espacios.

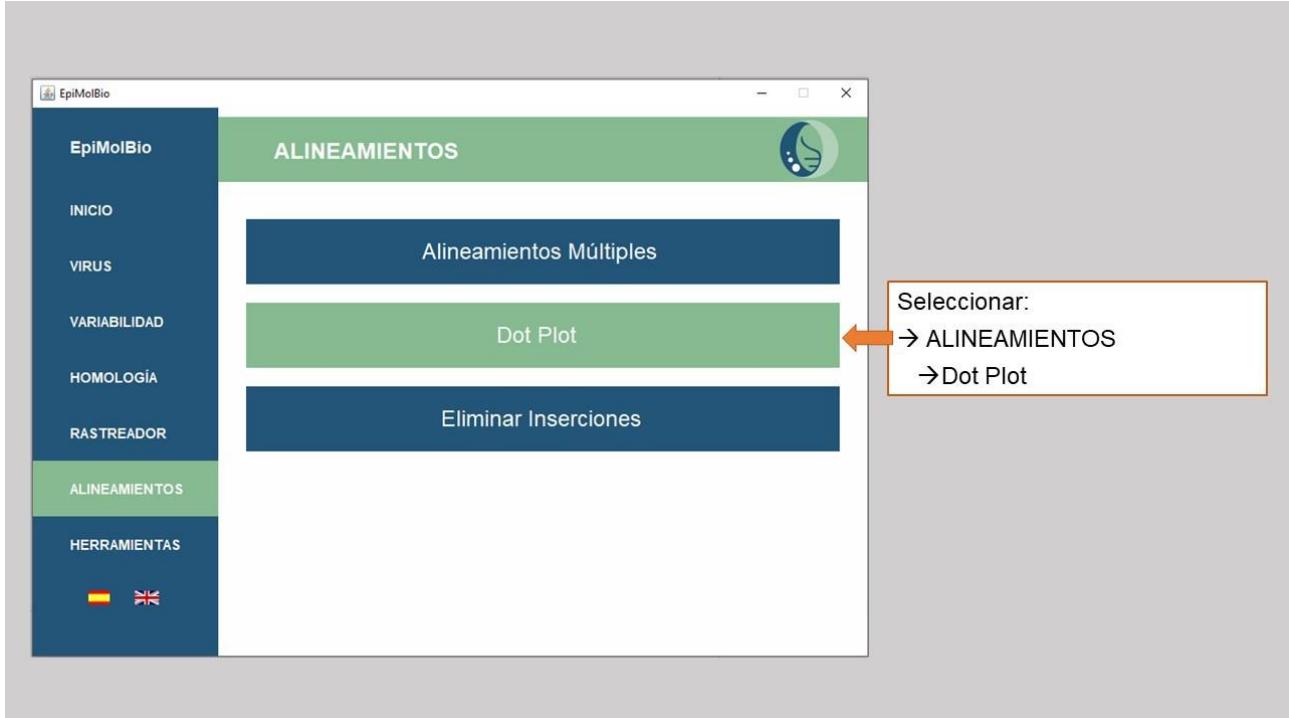
En el campo “**Secuencia 2**” introducir la segunda secuencia a comparar en nucleótidos o aminoácidos sin saltos de línea ni espacios.

Ambas secuencias deben estar en el mismo formato: las dos en nucleótidos o las dos en aminoácidos.

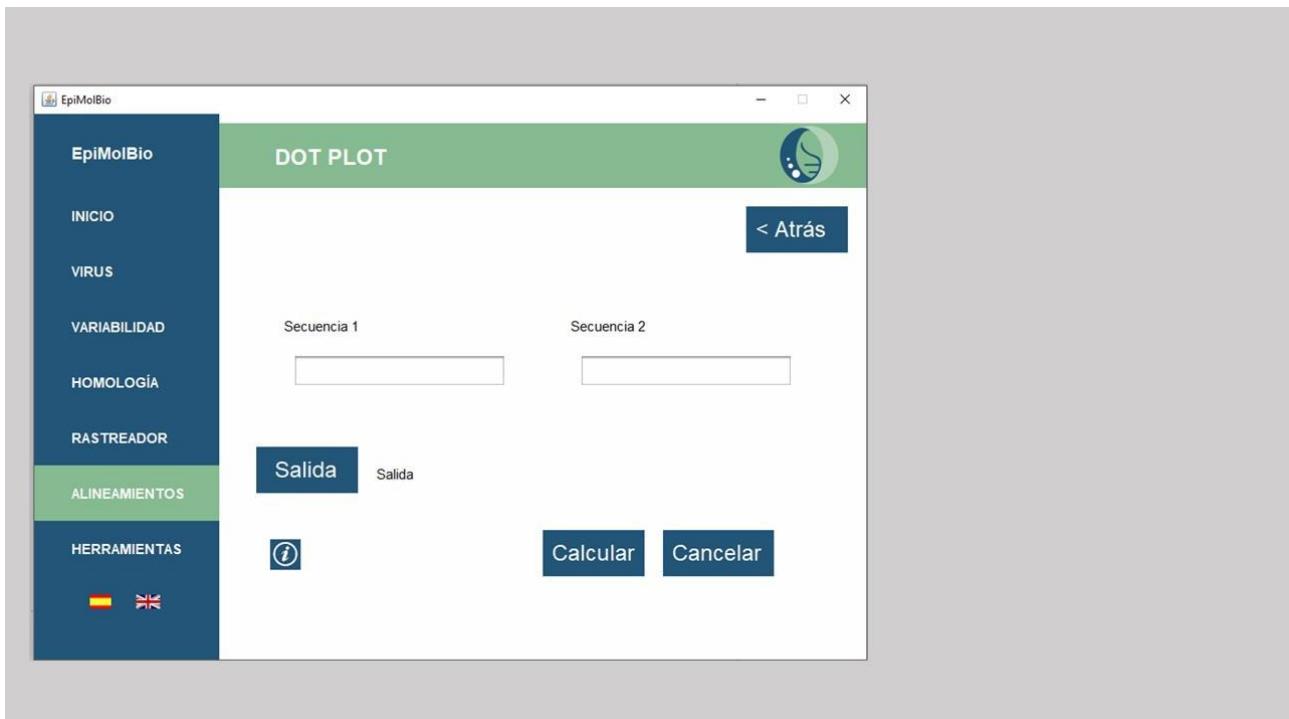
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo .html y nombrar los archivos escribiendo .html al final.

Paso a paso:

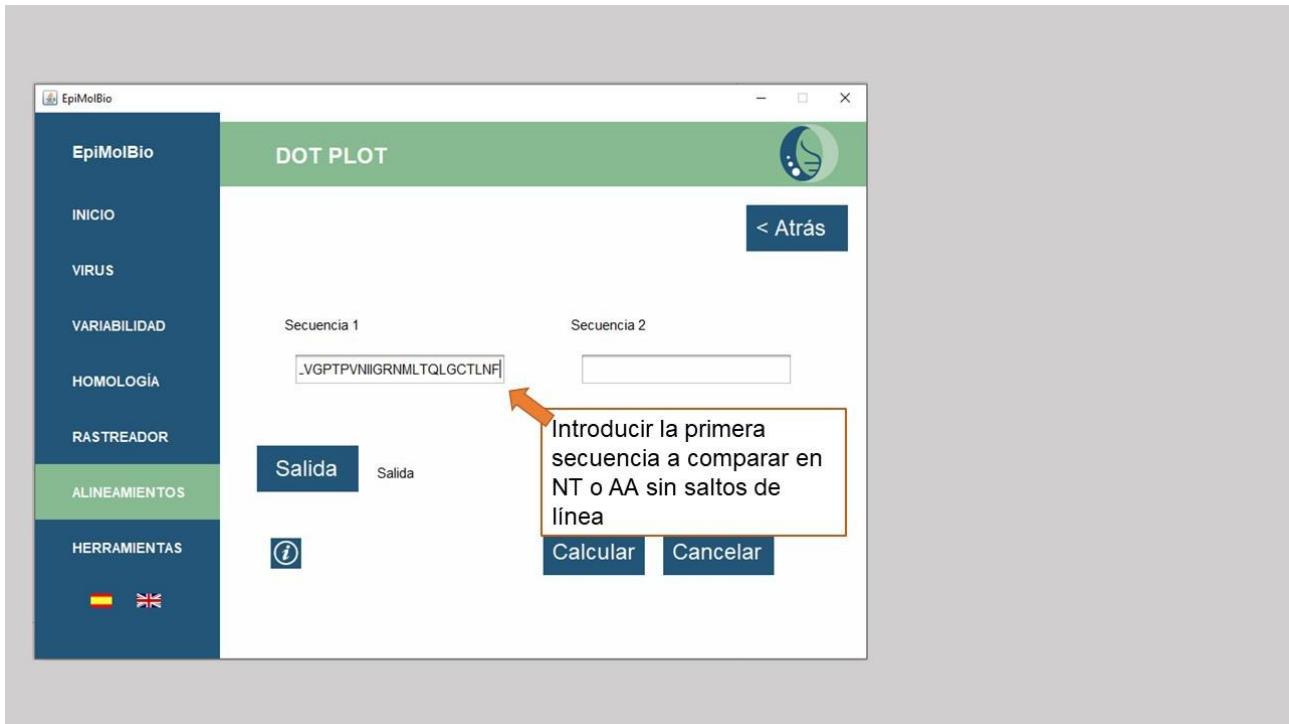
1)



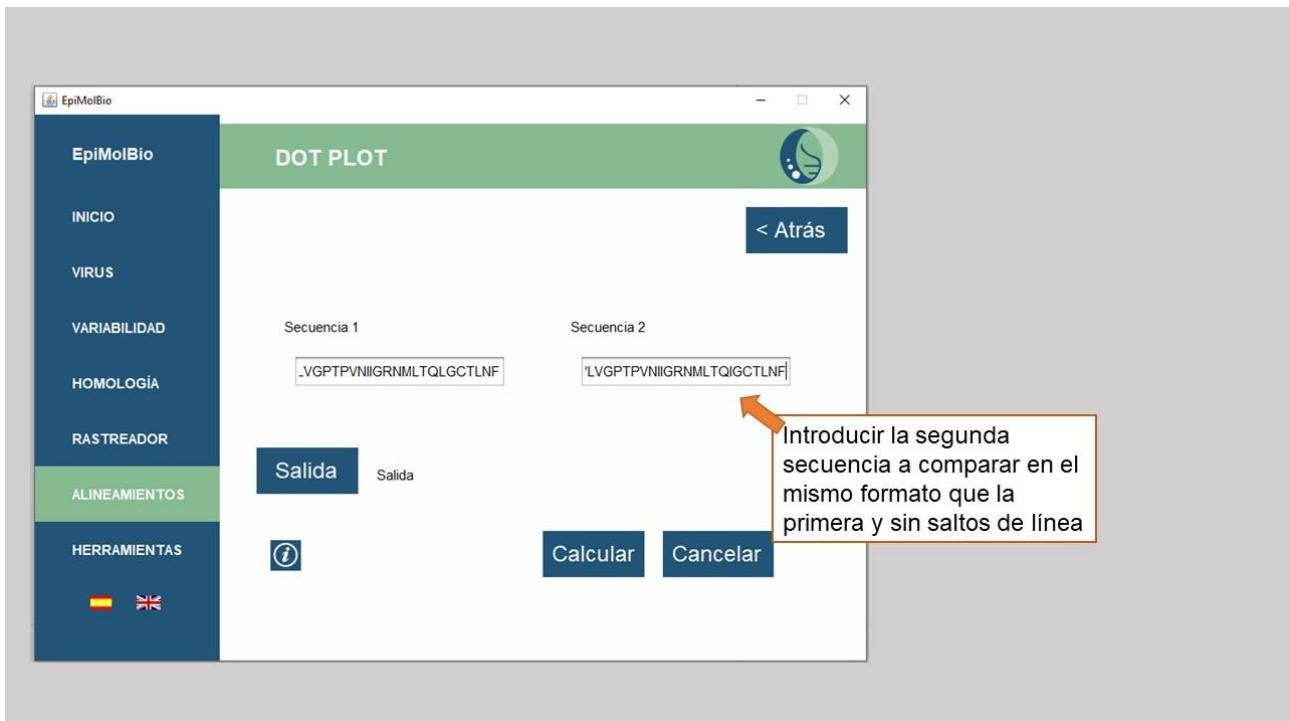
2)



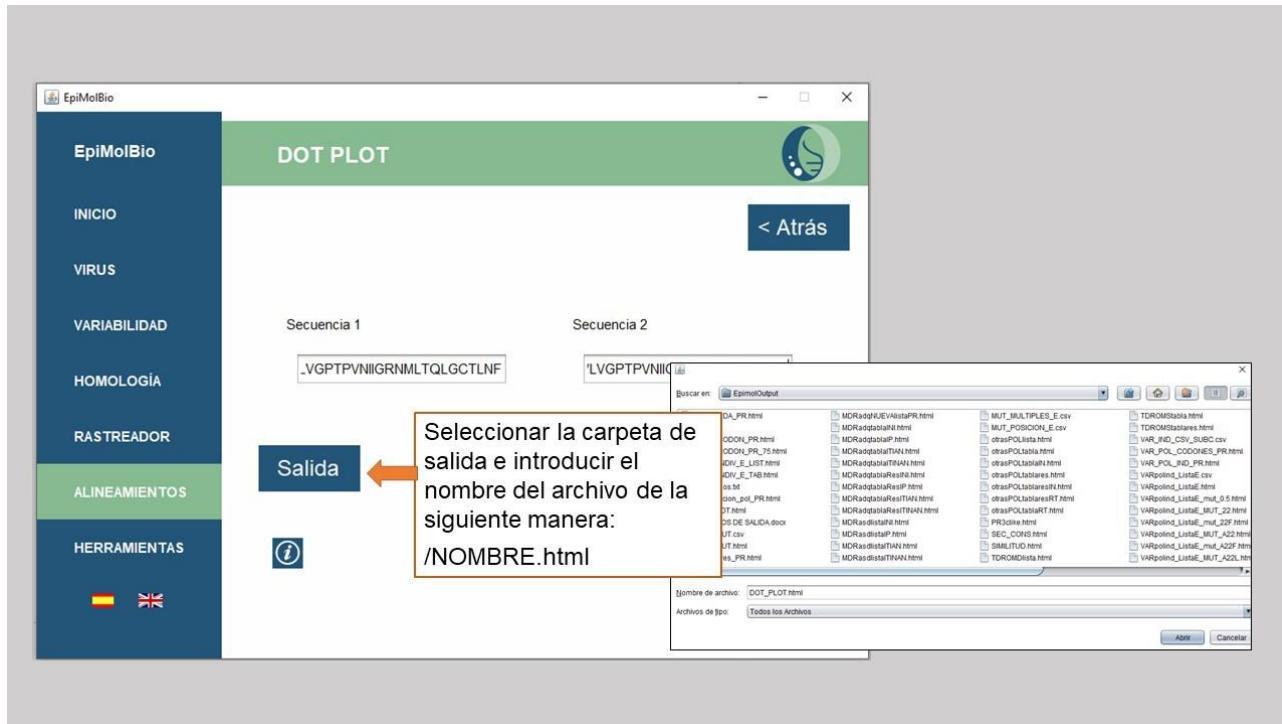
3)



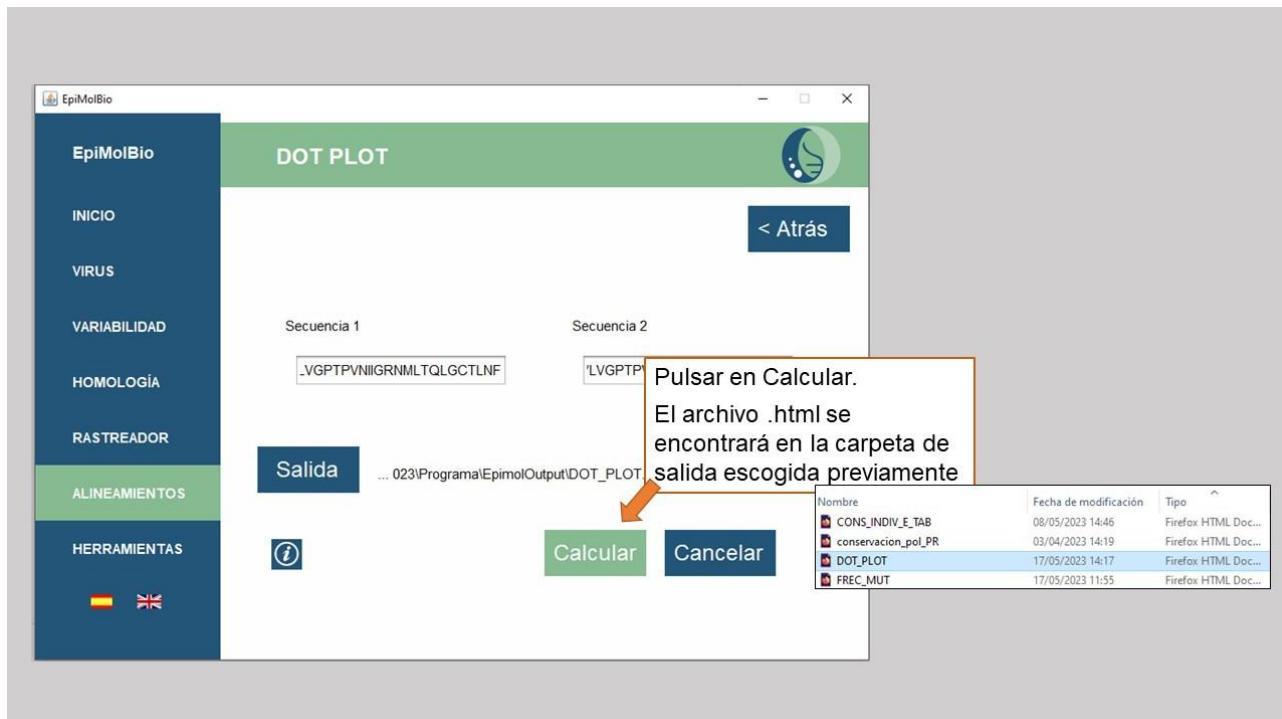
4)



5)



6)



V.3.ELIMINAR INSERCIÓNES

Esta función permite eliminar automáticamente las inserciones de las secuencias de un archivo con respecto a una referencia con gaps. Cuando las secuencias han sido previamente alineadas mediante un alineamiento múltiple, o descargadas de una base de datos que haga alineamientos múltiples con la referencia (esta debe ir en el archivo fasta que se descarga), tras el alineamiento, se generan gaps en la secuencia de referencia que corresponden a las inserciones. Esta función detecta los gaps en la secuencia de referencia y elimina dichas posiciones de las secuencias de entrada, eliminando las inserciones. El formato de salida es un archivo en formato .fasta con las secuencias resultantes sin inserciones.

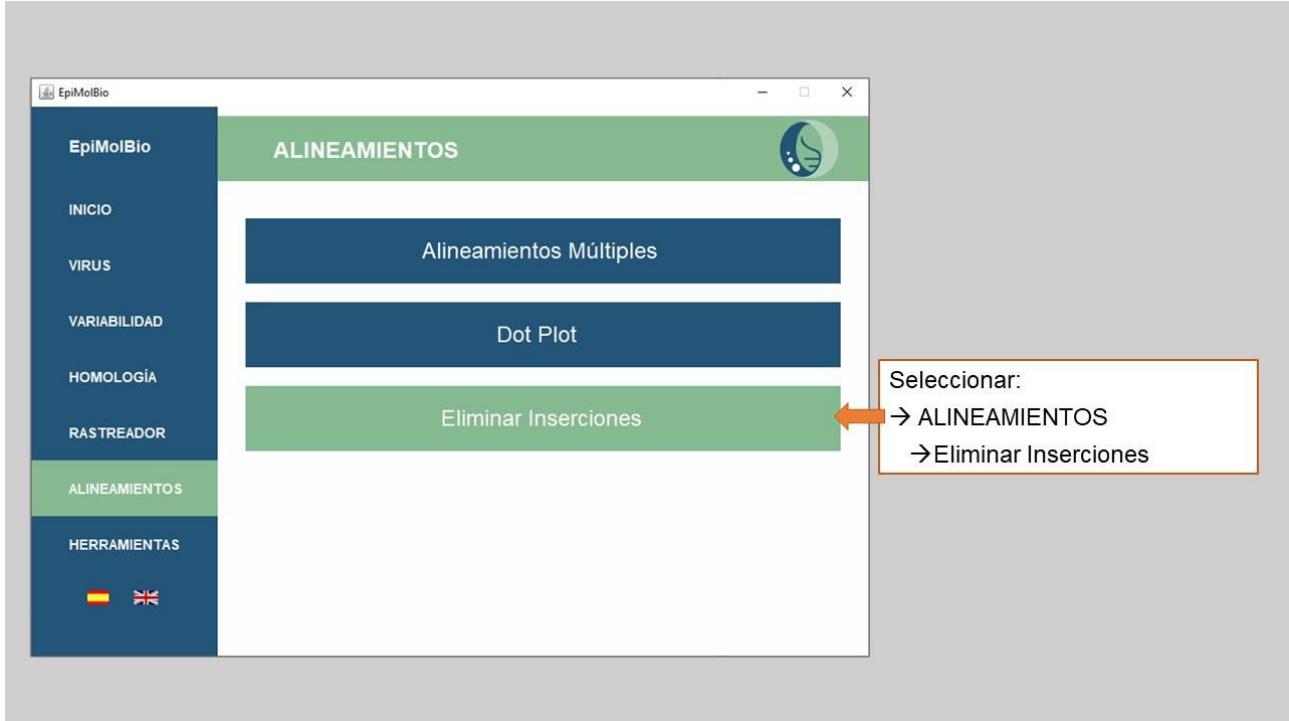
El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta con las secuencias alineadas en nucleótidos o aminoácidos.

En el campo “**Referencia**” introducir la secuencia de referencia con gaps, generada tras el alineamiento múltiple de las secuencias con la misma, sin saltos de línea ni espacios en aminoácidos o nucleótidos según los archivos de entrada.

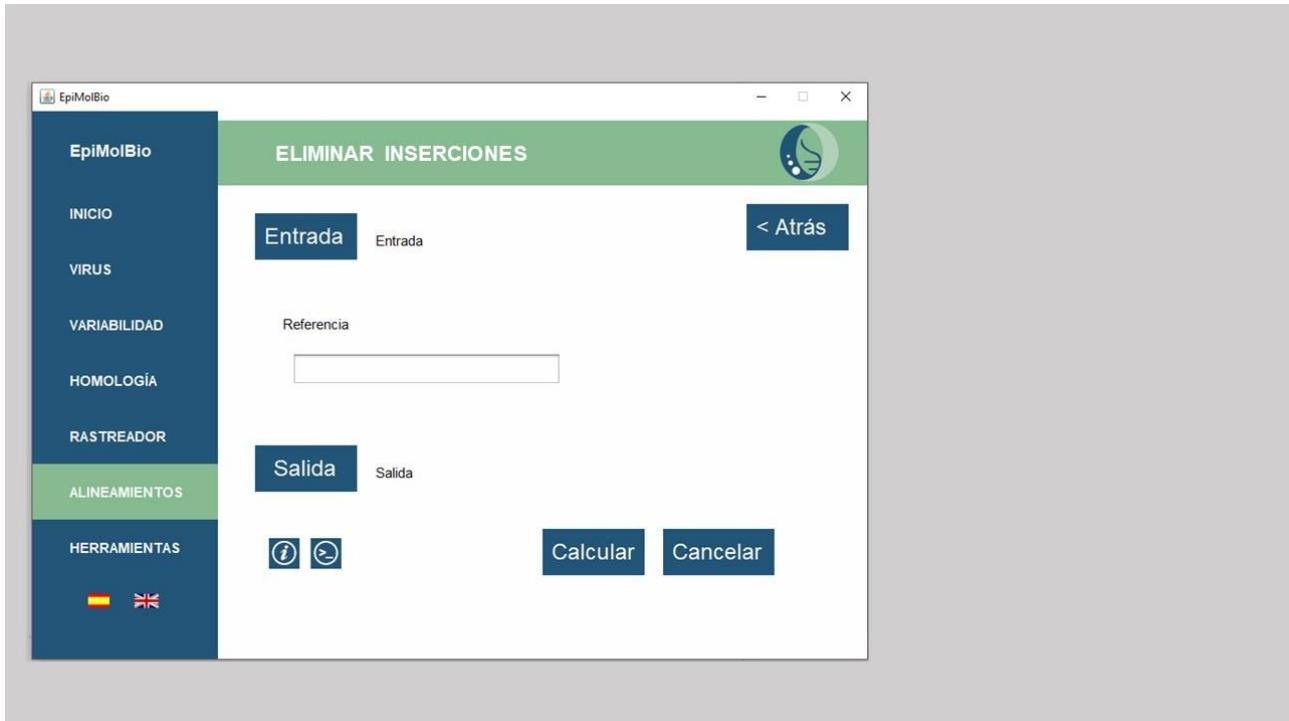
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezcan los archivos .fasta. Los archivos se nombran de forma automática de la siguiente forma: “Inserciones_Eliminadas_Nombre del archivo de entrada.fasta”.

Paso a paso:

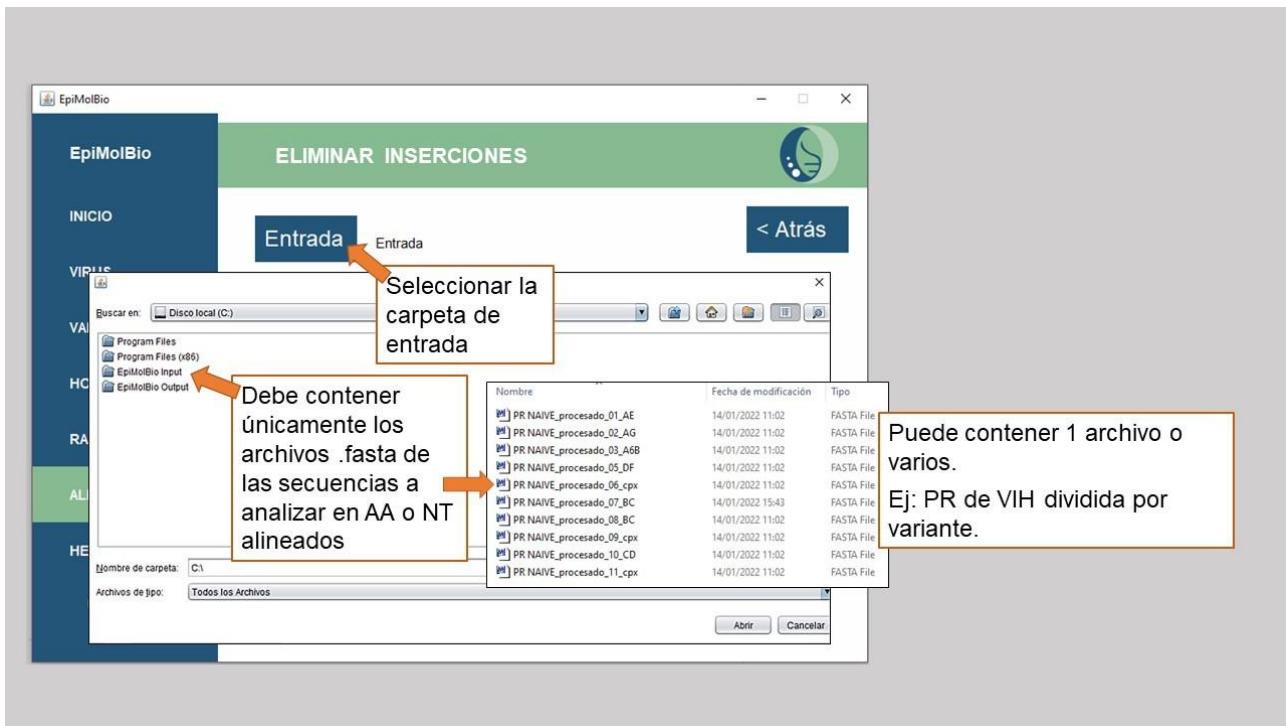
1)



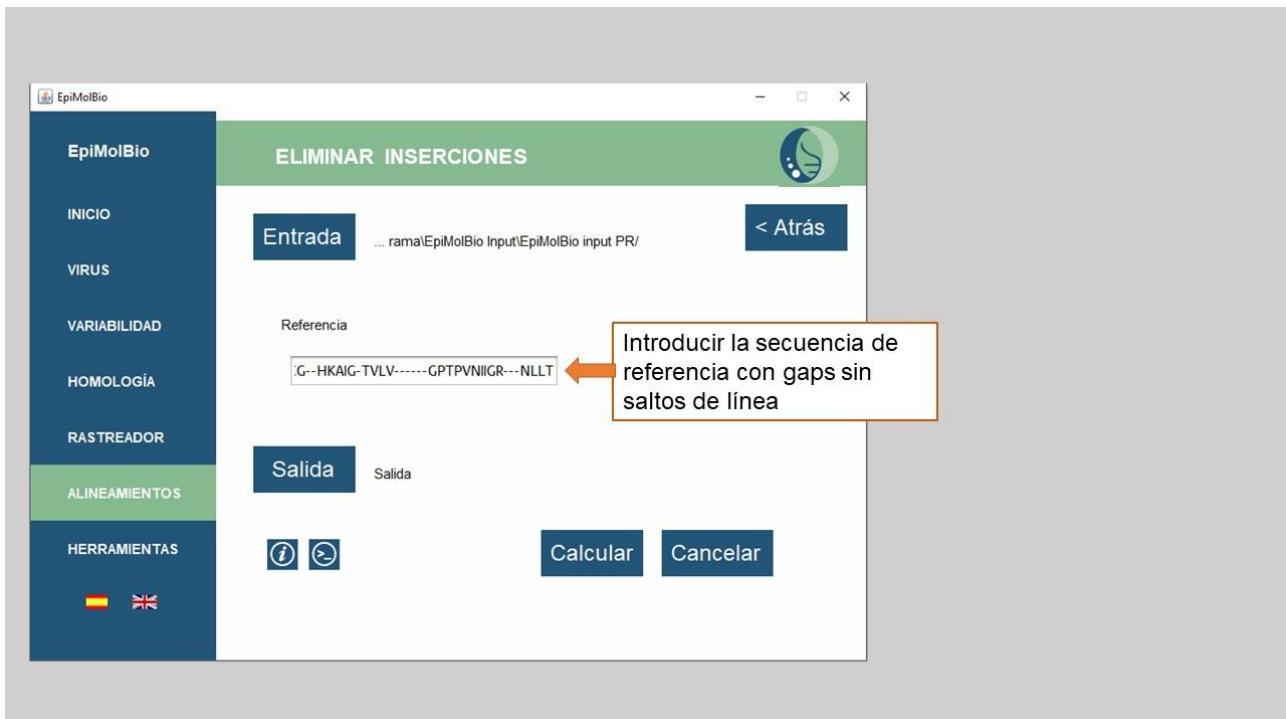
2)



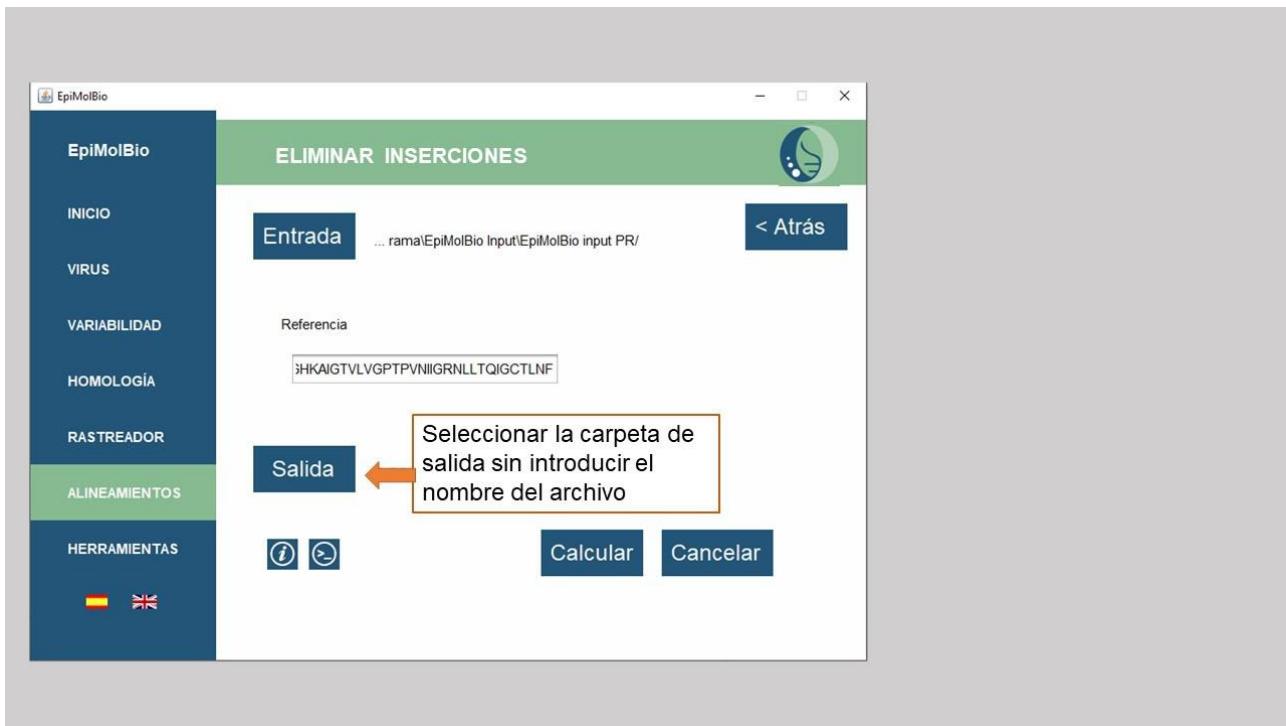
3)



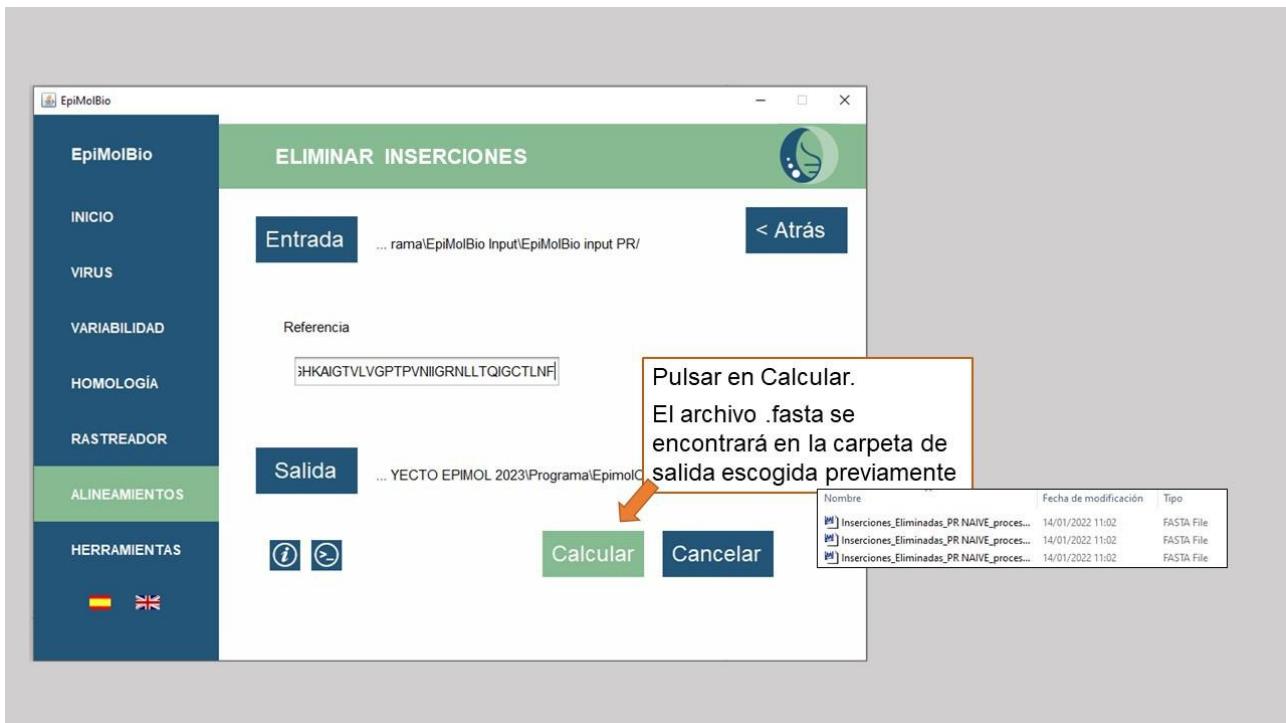
4)



5)



6)



VI. HERRAMIENTAS

En esta sección se encuentran varias **herramientas para modificar archivos y secuencias**, desde contar secuencias, editar encabezados o filtrar secuencias por su calidad, hasta crear un *workflow* de trabajo semiautomático con la herramienta “Programar Funciones”.

VI.1. EDICIÓN DE ARCHIVOS

VI.1.A) FUSIONAR ARCHIVOS

Esta herramienta **permite unir varios archivos .fasta en un solo archivo .fasta**.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiere fusionar. Los archivos pueden tener varias secuencias.

Marcar “**Fusionar Archivos**”.

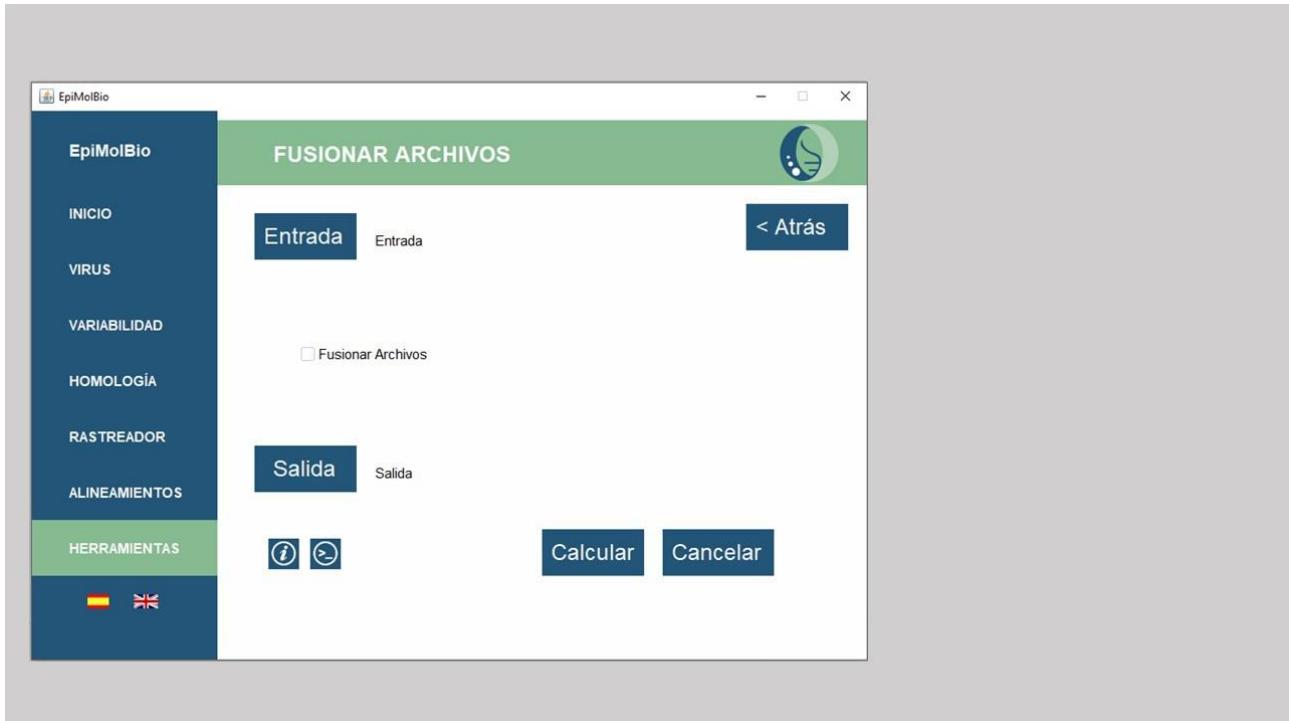
En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo .fasta y nombrarlo escribiendo .fasta al final. El formato de salida es un archivo en formato .fasta con todas las secuencias de los archivos de entrada.

Paso a paso:

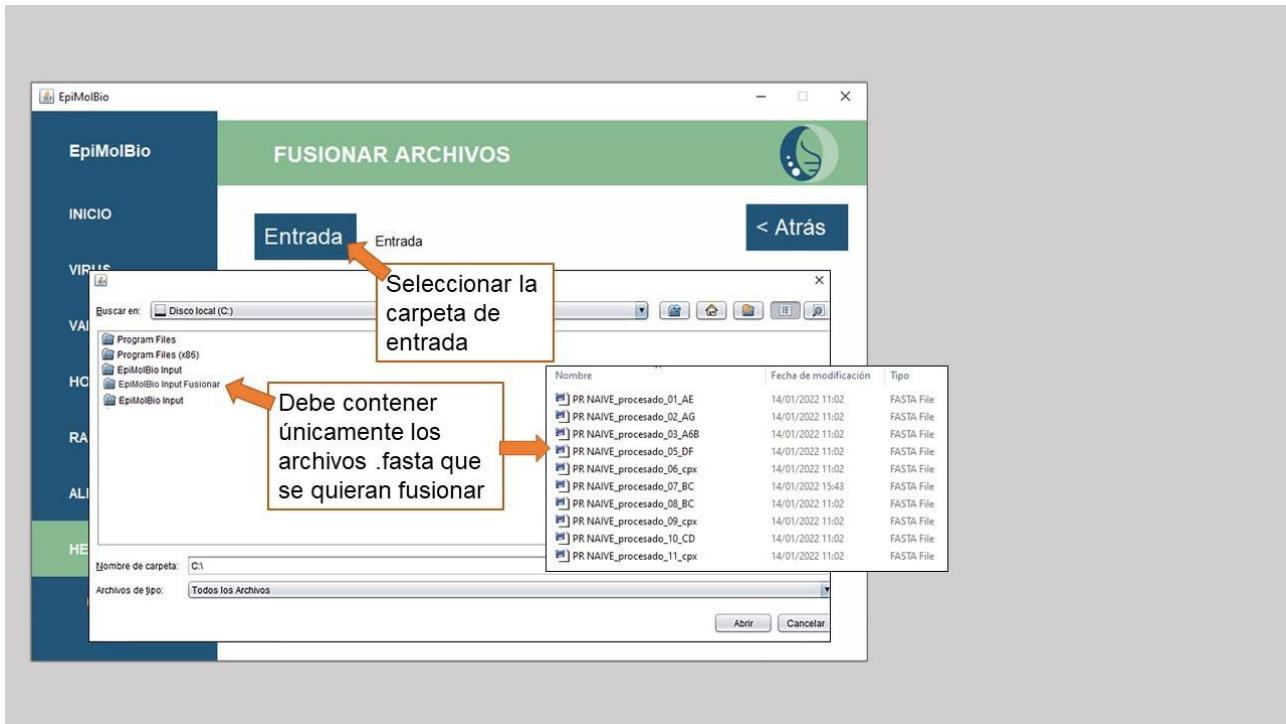
1)



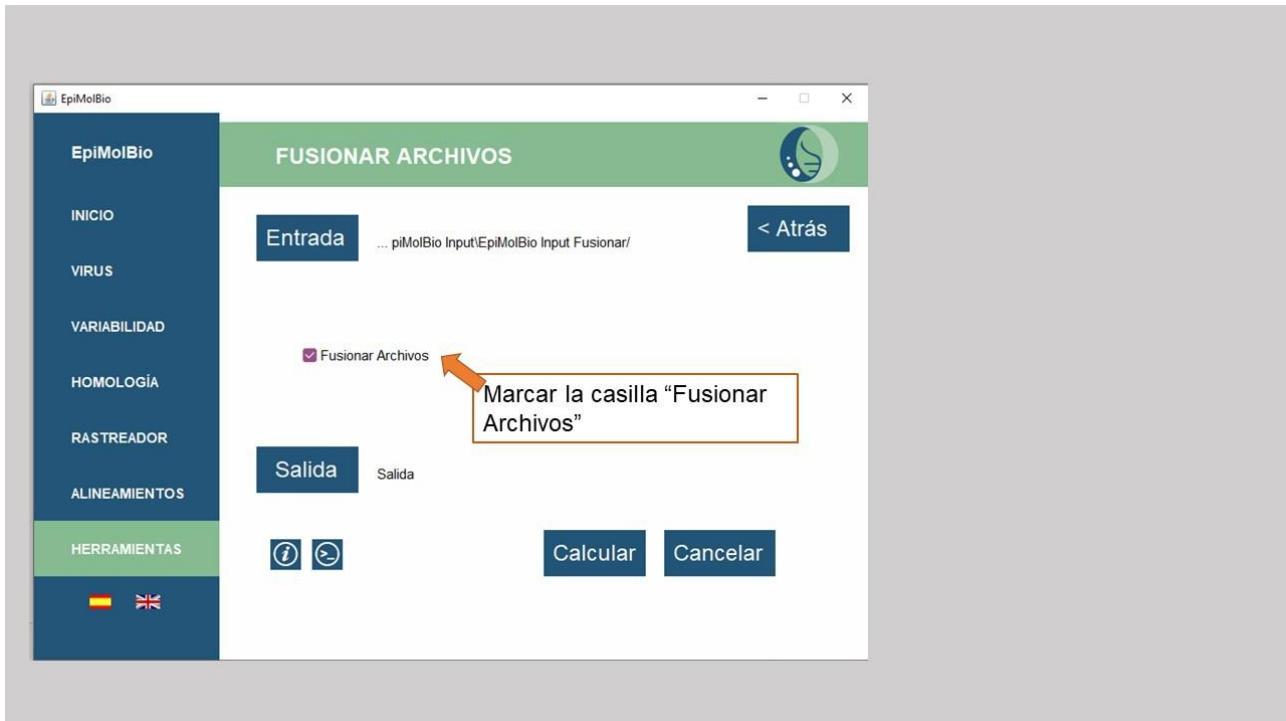
2)



3)



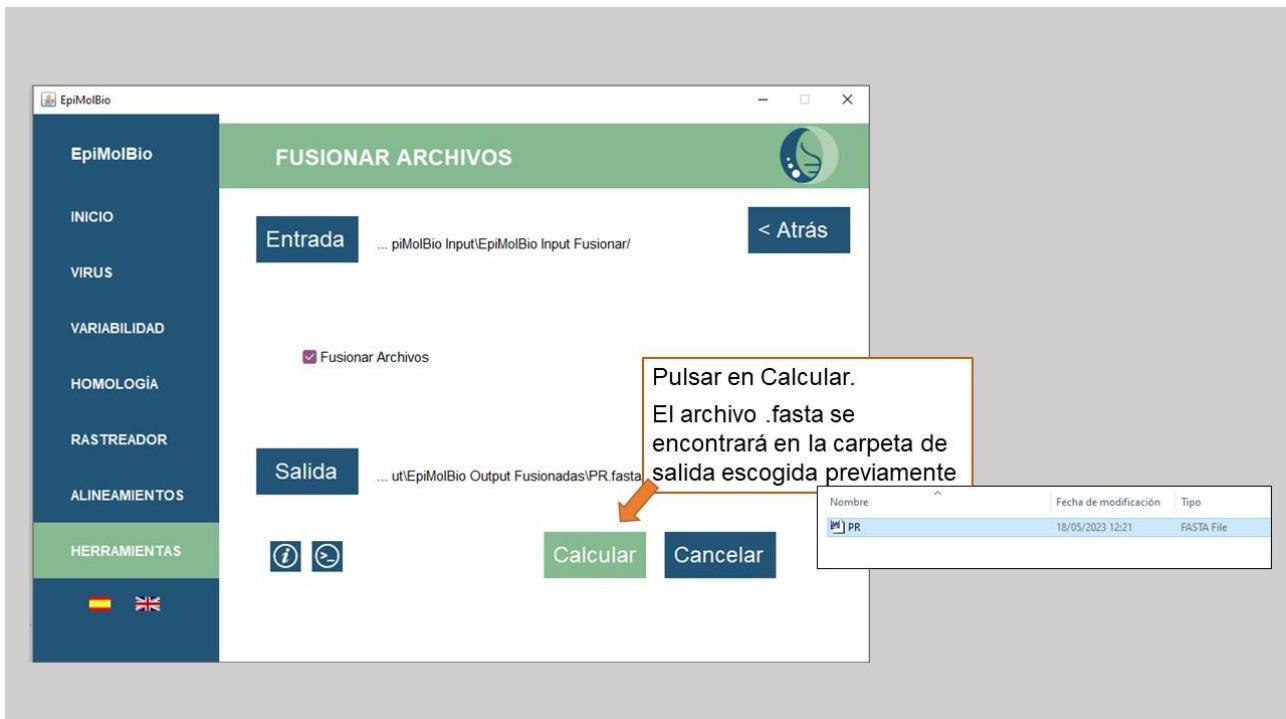
4)



5)



6)



VI.1.B) SECUENCIAS ÚNICAS

Esta herramienta **permite eliminar las secuencias repetidas de uno o varios archivos .fasta de entrada**, pudiendo escoger la frecuencia mínima con la que se quiere cribar las secuencias que se mantienen en el archivo de salida. Por ejemplo: tenemos 15 secuencias de la proteasa de la variante del VIH 06_cpx, 6 son iguales, aplicamos un cribado del 0.0 y obtenemos un archivo con 10 secuencias únicas, dejando una sola copia de aquellas que son idénticas a otras.

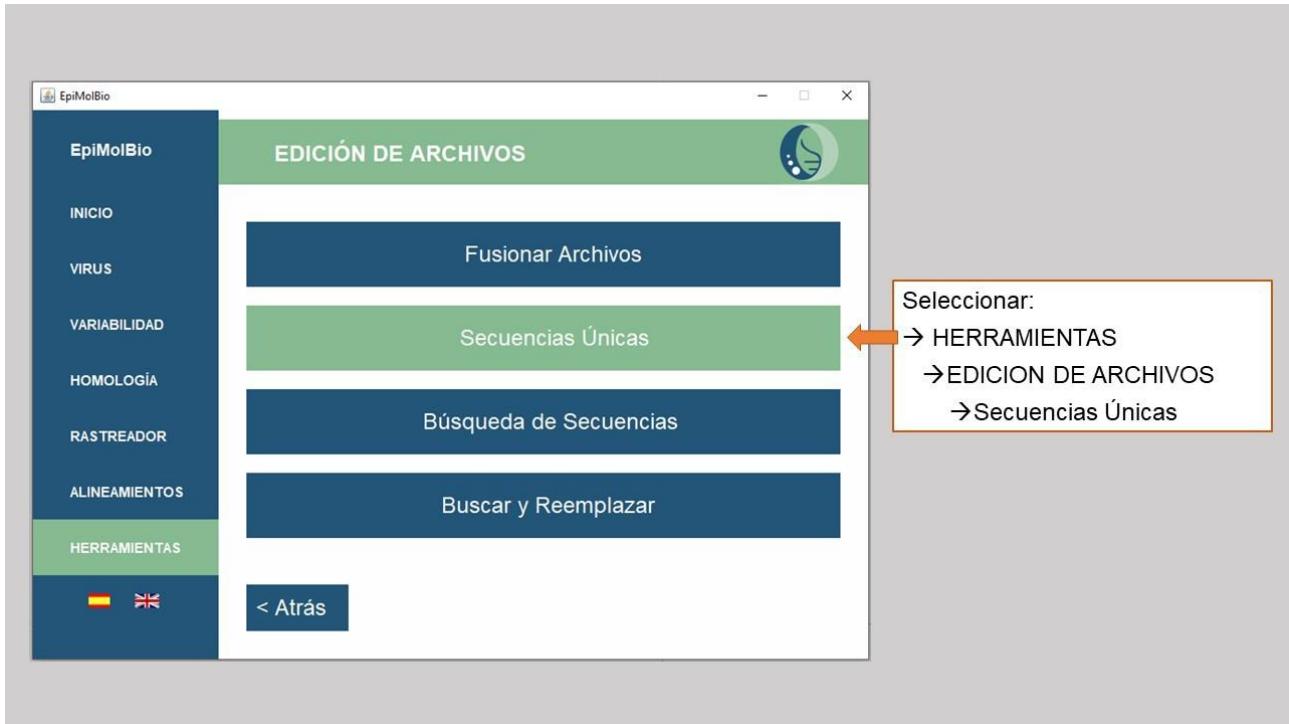
El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiere cribar.

En el campo “**Frecuencia Mínima**” introducir la frecuencia de cribado en números con un decimal. Por ejemplo, introducir 90.0 para realizar un cribado al 90% y que aparezcan solo las secuencias únicas con una representación mínima del 90%, o introducir 0.0 para que aparezcan todas las secuencias únicas.

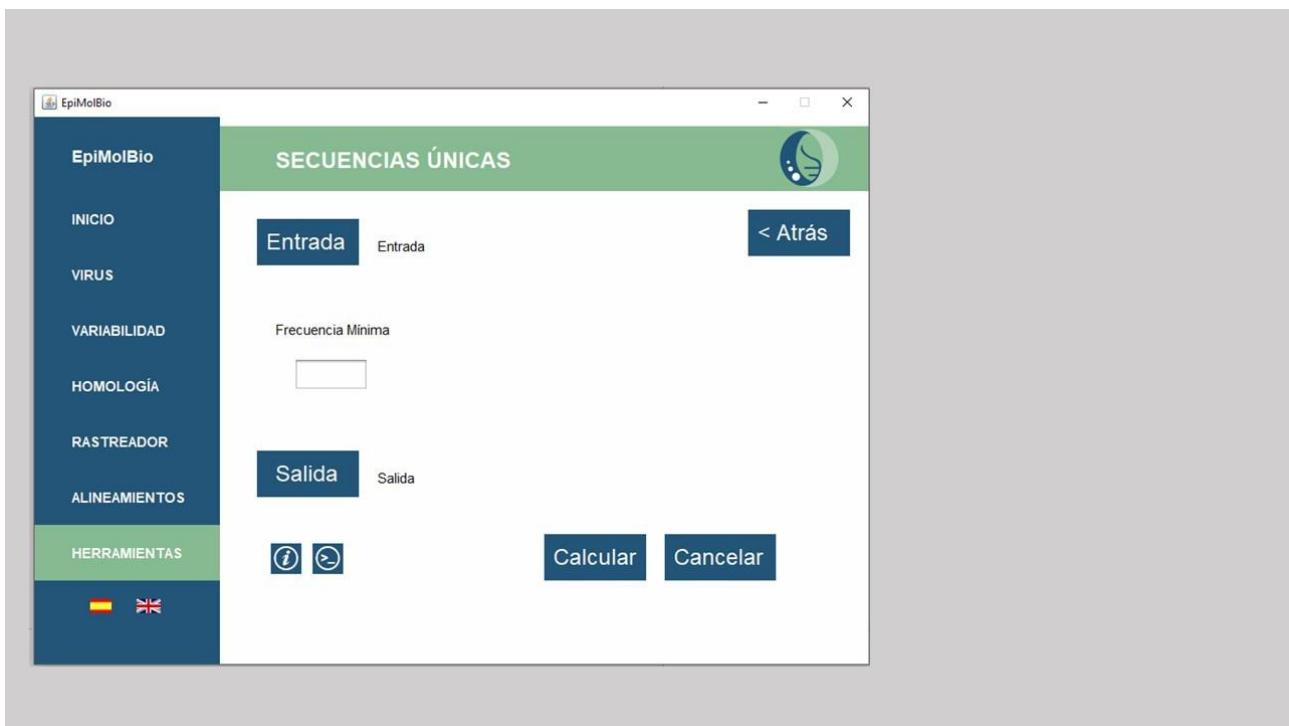
El formato de salida es un archivo en formato .fasta con las secuencias cribadas. En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo .fasta sin nombrarlo. Éstos se nombran de forma automática de la siguiente forma: “Únicas_Nombre del archivo de entrada.fasta”.

Paso a paso:

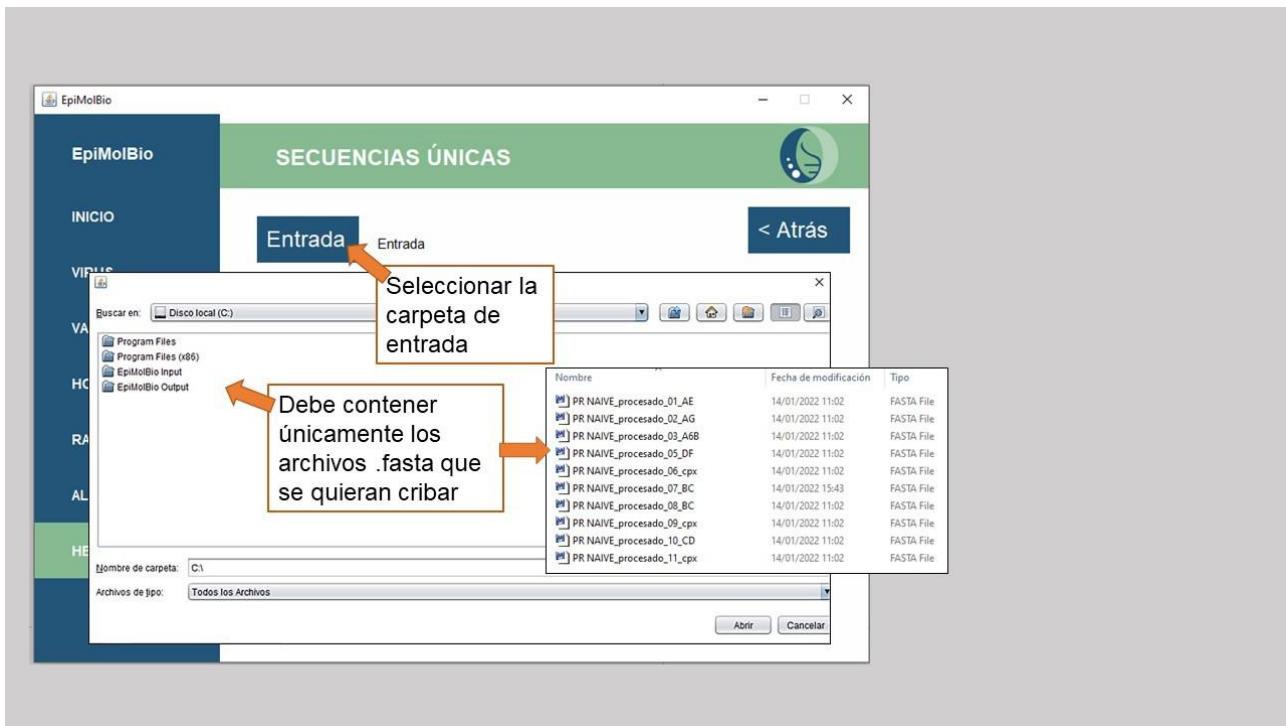
1)



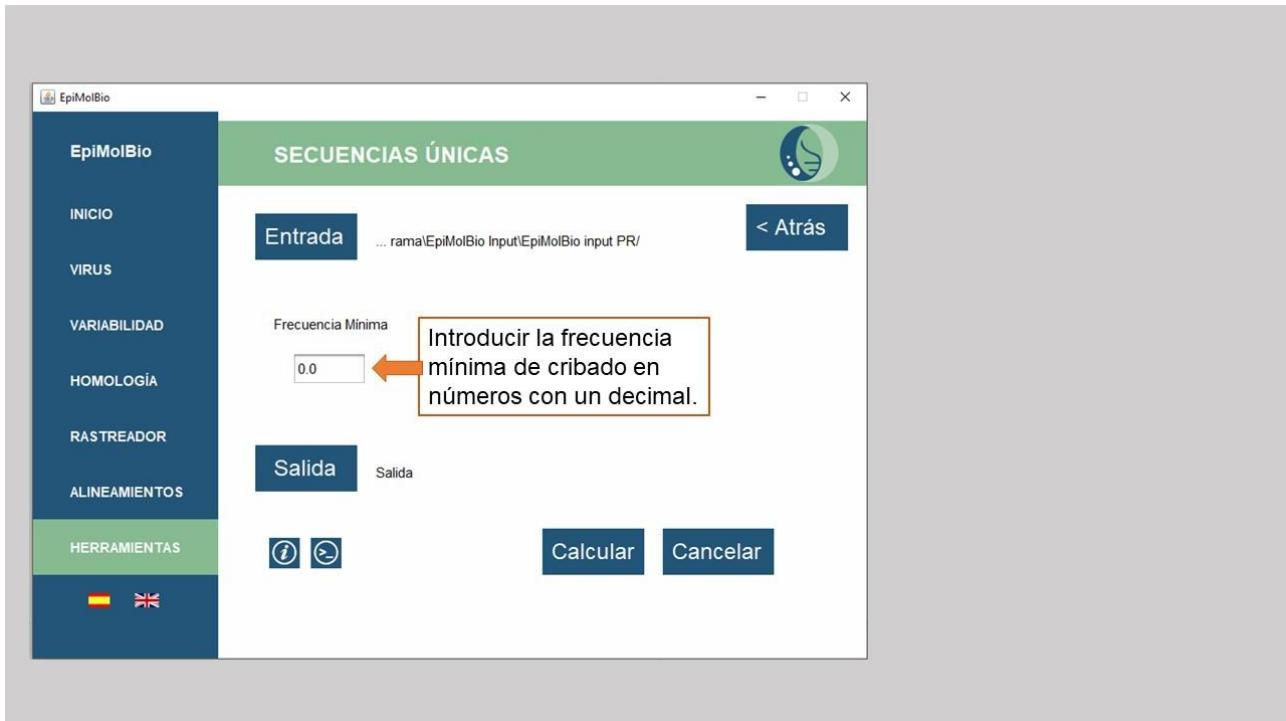
2)



3)



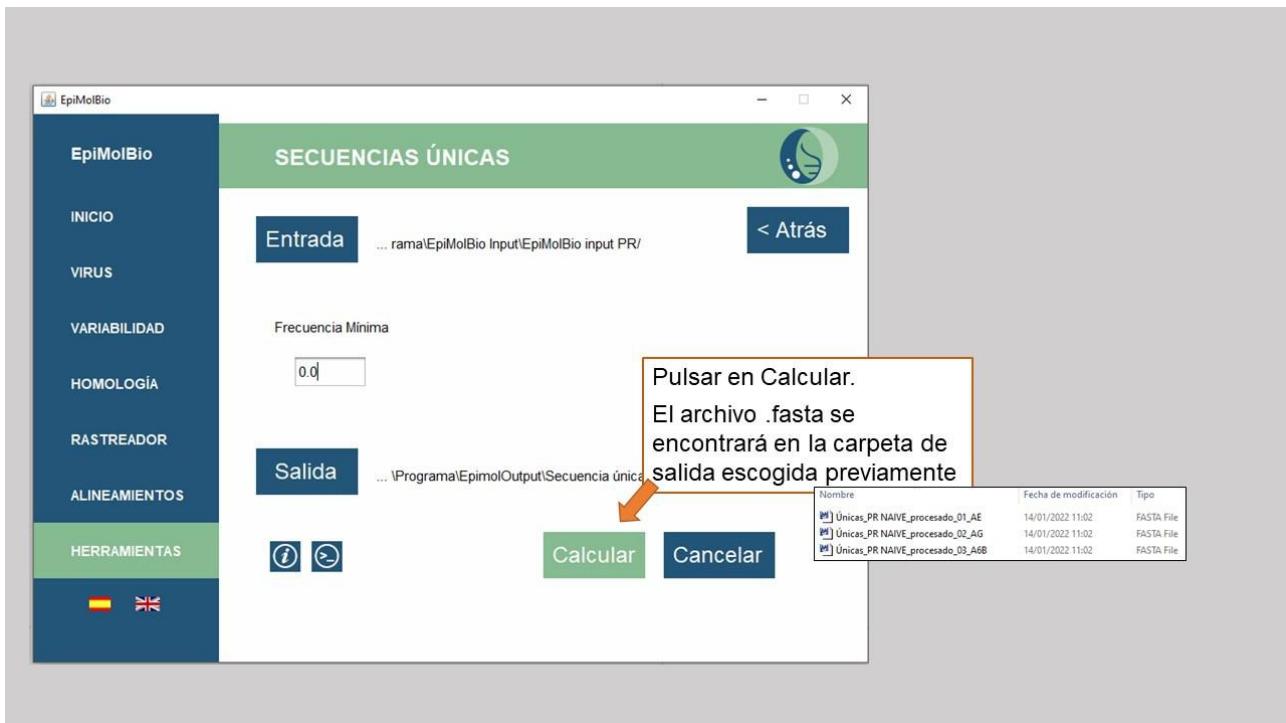
4)



5)



6)



VI.1.C) BÚSQUEDA DE SECUENCIAS

Esta función sirve para **cribar las secuencias de archivos .fasta que contengan una o varias mutaciones específicas**, pudiendo obtener como salida otro archivo .fasta o una tabla .csv.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiere cribar.

En el campo “**Formato**” escoger el tipo de salida entre FASTA (archivo .fasta) o CSV (archivo.csv).

El formato .csv consiste en una tabla que se puede abrir en Excel. En la parte superior aparece el archivo de entrada analizado. En la primera columna se muestran los encabezados de las secuencias resultantes del análisis y en la segunda, sus secuencias. Las filas vacías indican que no se han encontrado las mutaciones escogidas dentro de ninguna de las secuencias del archivo de entrada indicado en la primera columna.

Ejemplo de formato de salida .csv del análisis Búsqueda de Secuencias:

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PR_01_AE.fasta | | | | | | | | | | |
| 2 | >01_AE.VN.2 PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQISIECGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | | | | | | | | | | |
| 3 | >01_AE.TH.2I PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQILIECGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | | | | | | | | | | |
| 4 | >01_AE.CN.2 PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQILIECGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF | | | | | | | | | | |
| 5 | >01_AE.CN.2 PQITLWQRPLVTIKIEGQLKEALLDTGADDTVLEDINLPGKLKPVIGGIGGFIKVRQYDQILIELCGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | | |
| 6 | >01_AE.CN.2 PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEDINLPGKWKPKVIGGIGGFIKVRQYDQIPIECGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | | |
| 7 | >01_AE.CN.2 PQITLWQRPCVTIKIGAELKEALLDTGADDPVLEDINLPGKWKPKVIGGIGGFIRVRHYDRVVIGICGRKAVRTVLVRPTPVNIKRNMFSHGLFALNF | | | | | | | | | | |
| 8 | >01_AE.CN.2 PQITLWQRPLVTVKIGDQLREALLDTGADDTVLEEINLPGKWKPKVIGGIGGFIKVRQYDQISIECGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | | |
| 9 | >01_AE.CN.2 PQITLWQRPLVTVKIGDQLREALLDTGADDTVLEEINLPGKWKPKVIGGIGGFIKVRQYDQISIECGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF | | | | | | | | | | |

El formato .fasta genera un archivo .fasta por cada archivo de entrada introducido. Aquellos en los que no se han detectado las mutaciones escogidas no aparecerán.

Ejemplo de formato de salida .fasta del análisis Búsqueda de Secuencias:

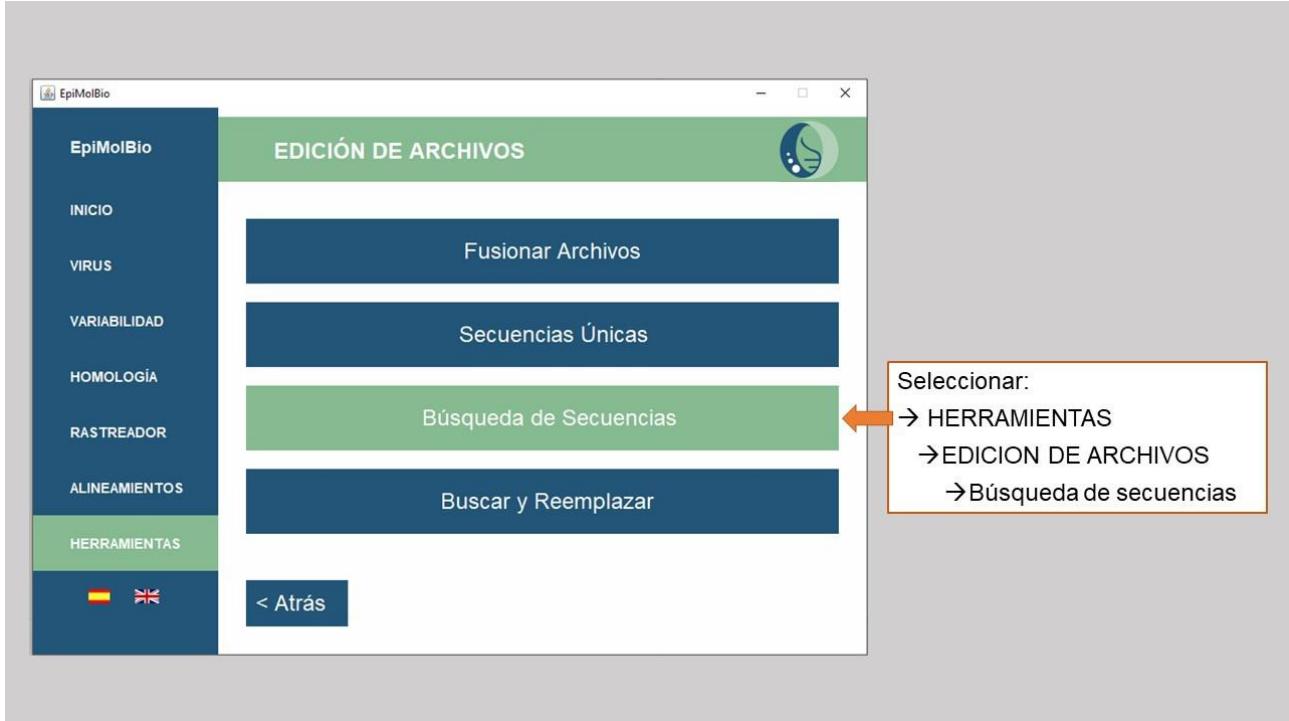
| Nombre | Fecha de modificación | Tipo |
|--|-----------------------|------------|
| [+] Búsqueda_IN_NAIVE_procesado_01_AE | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_02_AG | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_06_cpx | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_07_BC | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_08_BC | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_14_BG | 01/04/2020 15:43 | FASTA File |
| [+] Búsqueda_IN_NAIVE_procesado_16_A2D | 01/04/2020 15:43 | FASTA File |

En el campo “**Mutaciones**” escribir una mutación o varias separadas por “,” y sin espacios (ej.: M50I,L74I).

En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo con los resultados. Si se escoge el formato de salida .csv habrá que nombrar el archivo escribiendo .csv al final. Si se escoge el formato de salida .fasta no habrá que nombrar el archivo ya que se nombra automáticamente de la siguiente forma: “Búsqueda_Nombre del archivo de entrada.fasta”.

Paso a paso:

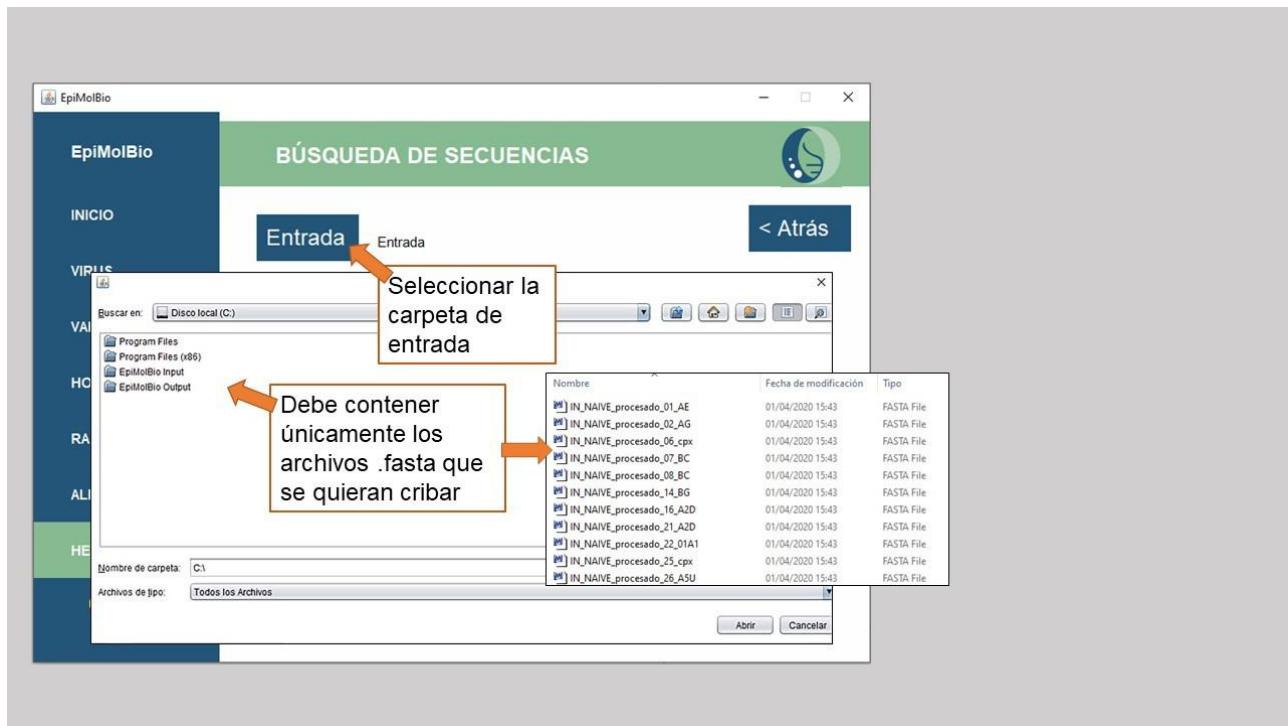
1)



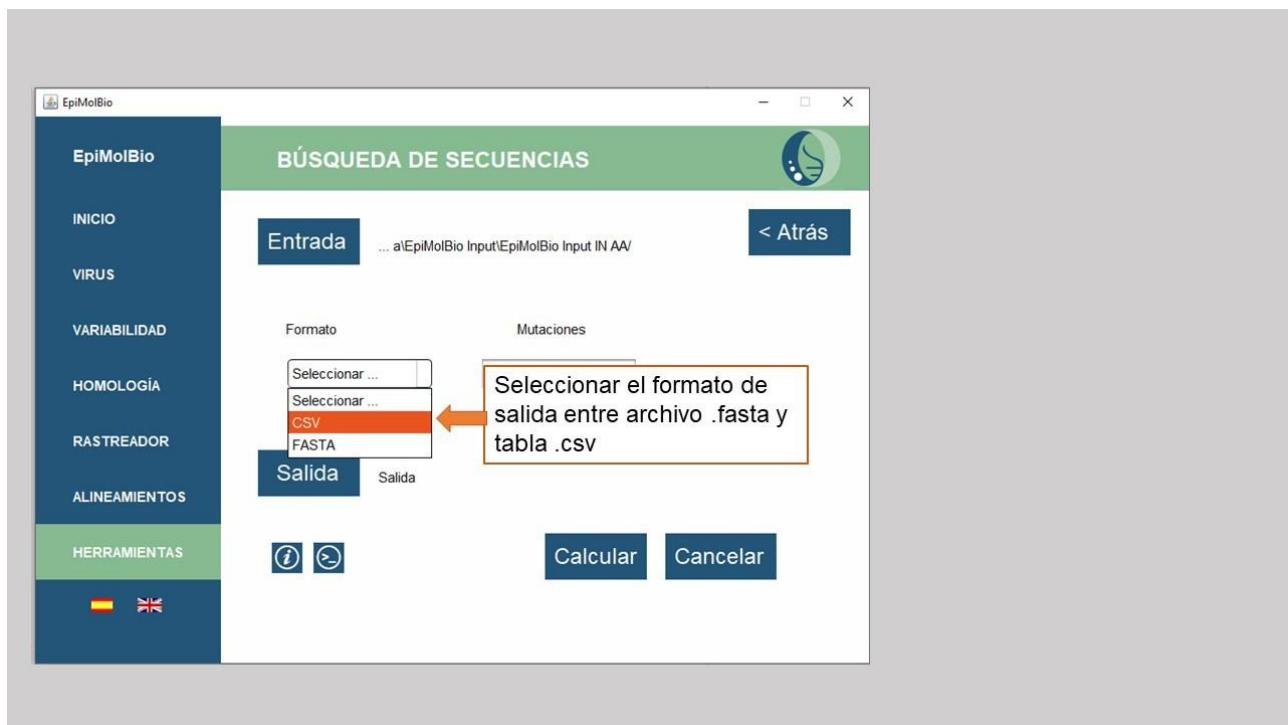
2)



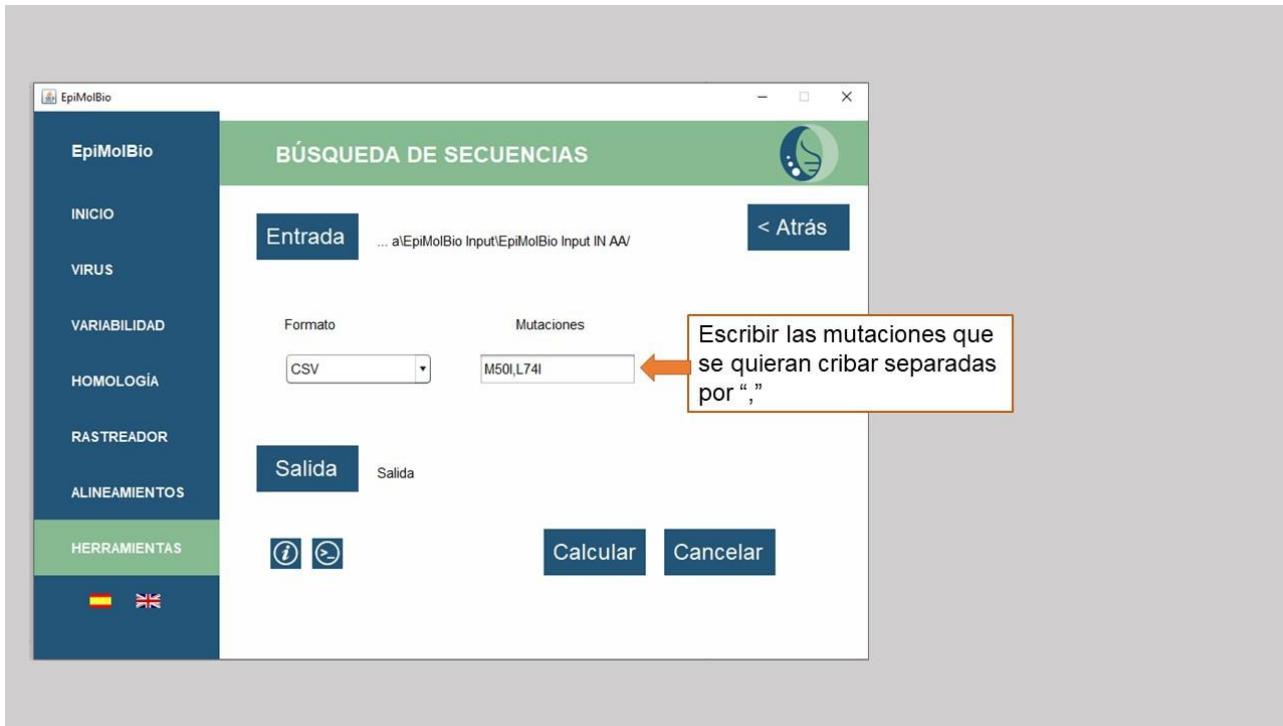
3)



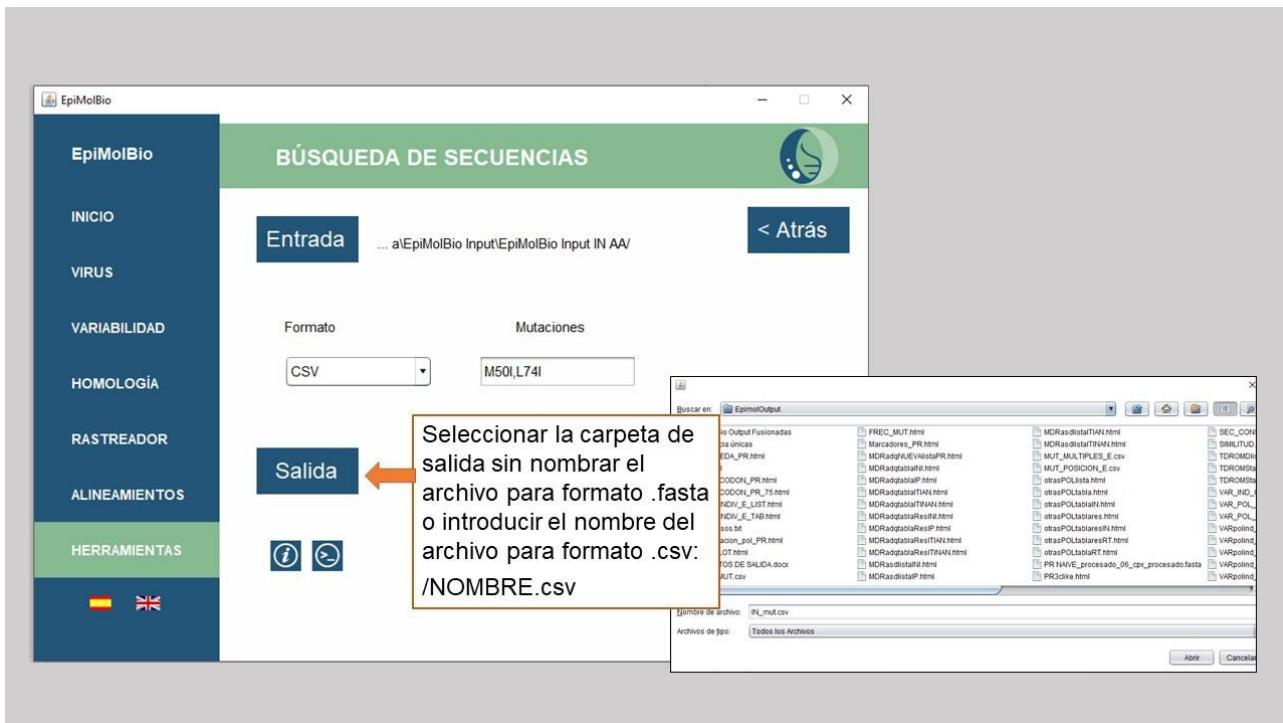
4)



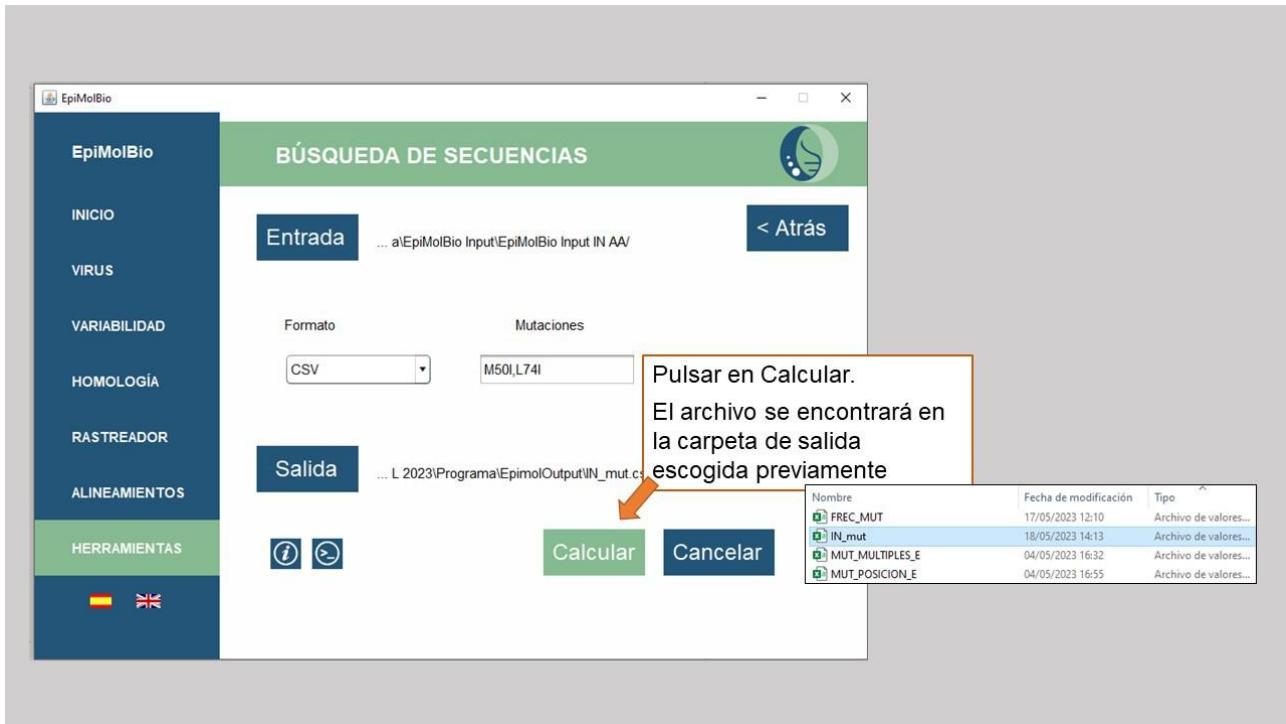
5)



6)



7)



VI.1.D) BUSCAR Y REEMPLAZAR

Esta herramienta **permite reemplazar en uno o varios archivos “.fasta” una serie de caracteres por otros, tanto en el encabezado como en la propia secuencia genética.** Por ejemplo: cambiar en el encabezado de los archivos .fasta un “-” por “/” para poder usar posteriormente la función de filtrado por encabezado, o reemplazar “N” por “?” para realizar un análisis en secuencias de nucleótidos con una función para la que EpiMolBio sólo permita secuencias en aminoácidos como entrada, como es el caso de “Variabilidad, Polimorfismos, Individual”.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiera modificar.

En el campo “**Buscar**” introducir los caracteres a reemplazar (ej.: _).

En el campo “**Reemplazar**” introducir los caracteres nuevos que reemplazarán los anteriores (ej.: /)

Escoger “**Secuencia**” si se van a reemplazar caracteres dentro de la secuencia o “**Encabezado**” si se van a reemplazar dentro del encabezado.

En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo .fasta sin nombrarlo. Éstos se nombran de forma automática de la siguiente forma: “Reemplazar_Nombre del archivo de entrada.fasta”.

Ejemplo de archivo original y archivo modificado de la herramienta Buscar y Reemplazar:

```
>HCOV/19/SPAIN/AS/232252631/2022.EPI|ISL|8818639.2022/01/05
MYSFVSEETGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPS
FYVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232253923/2022.EPI|ISL|8818658.2022/01/05
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232253886/2022.EPI|ISL|8818657.2022/01/05
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232260023/2022.EPI|ISL|8818668.2022/01/07
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
```

```
>HCOV/19/SPAIN/AS/232252631/2022.EPI|ISL|8818639.2022/01/05
MYSFVSEETGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPS
FYVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232253923/2022.EPI|ISL|8818658.2022/01/05
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232253886/2022.EPI|ISL|8818657.2022/01/05
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
>HCOV/19/SPAIN/AS/232260023/2022.EPI|ISL|8818668.2022/01/07
MYSFVSEEIGTLIVNSVLLFLAFVVFLVTLAILTALRLCAYCCNIVNVSLVKPSF
YVYSRVKNLNNSSRVPDLLV
```

Paso a paso:

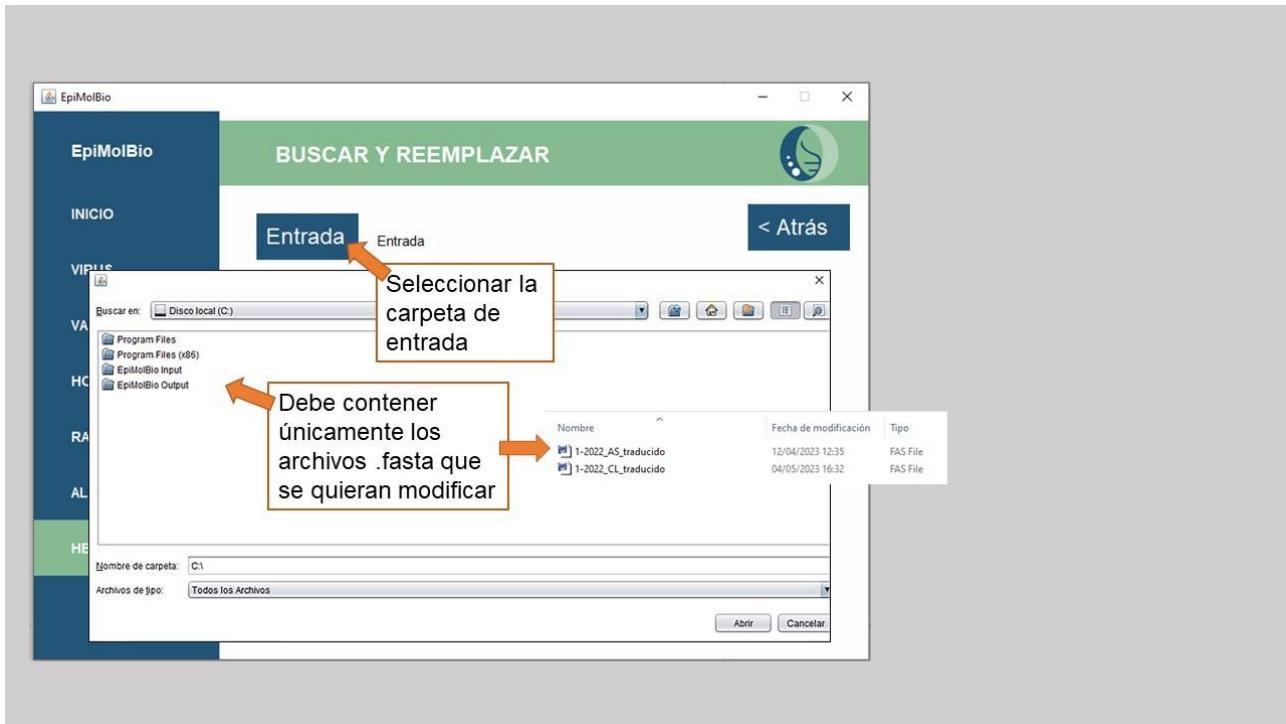
1)



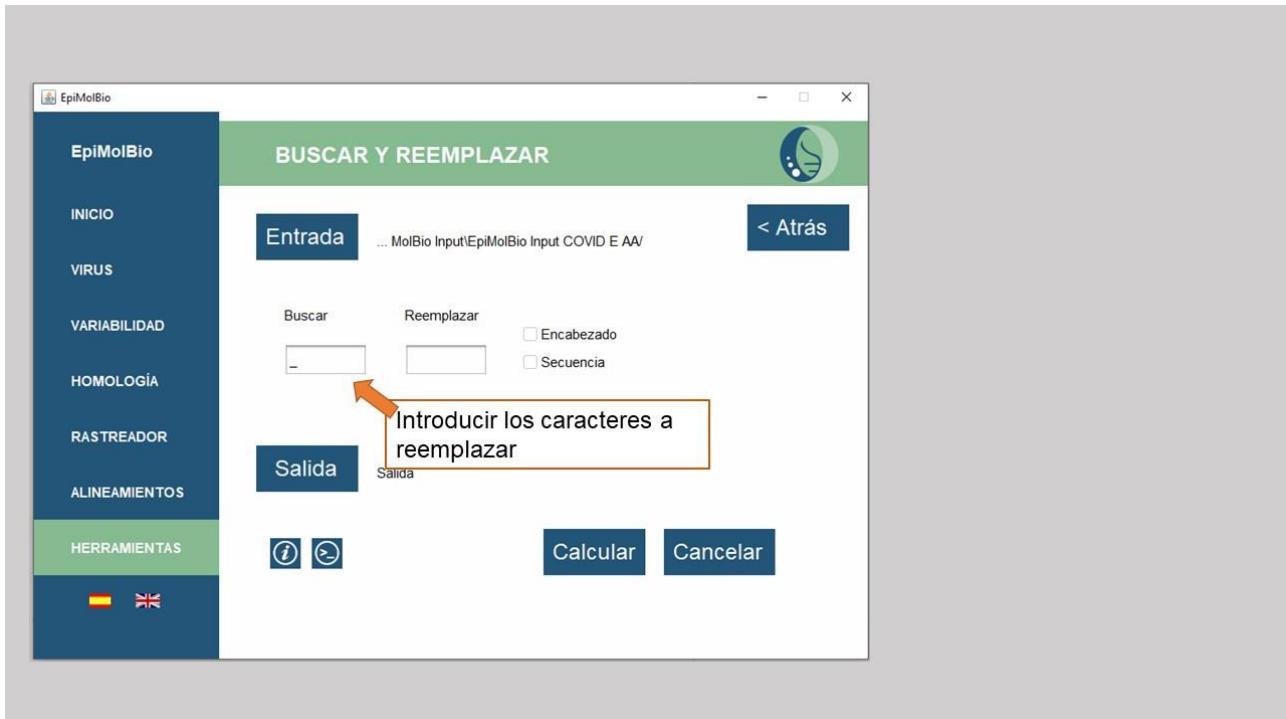
2)



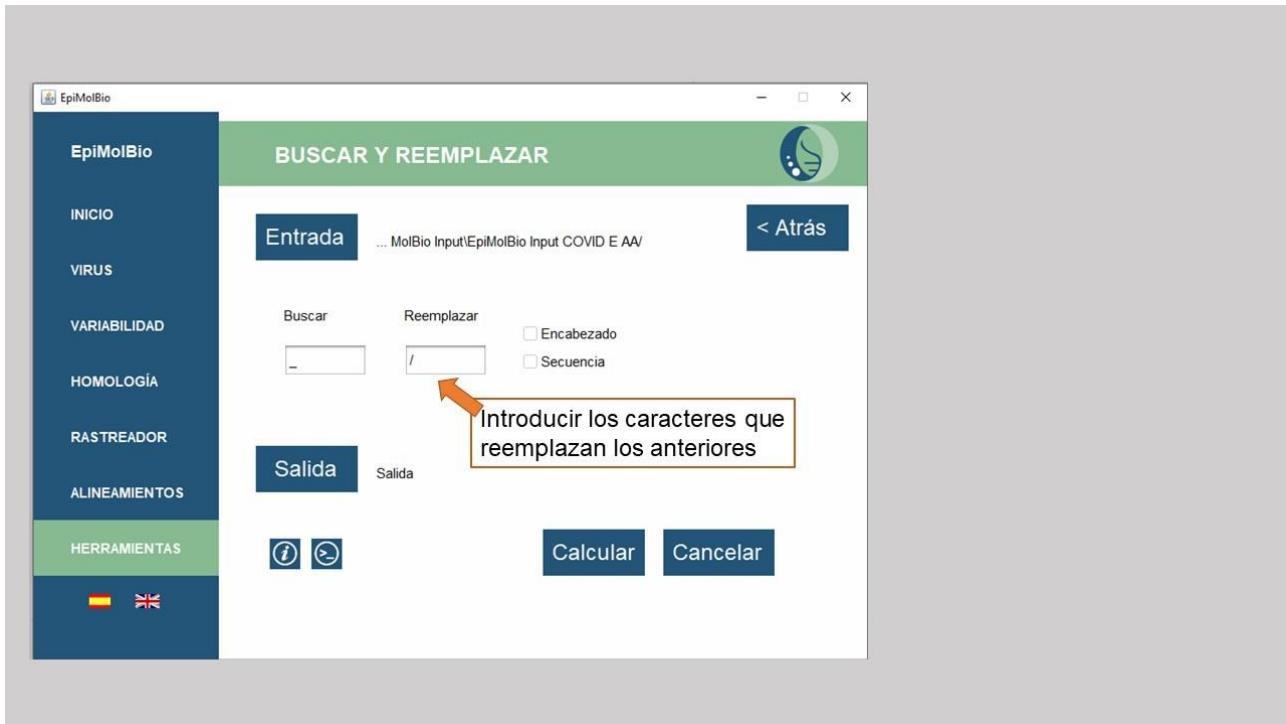
3)



4)



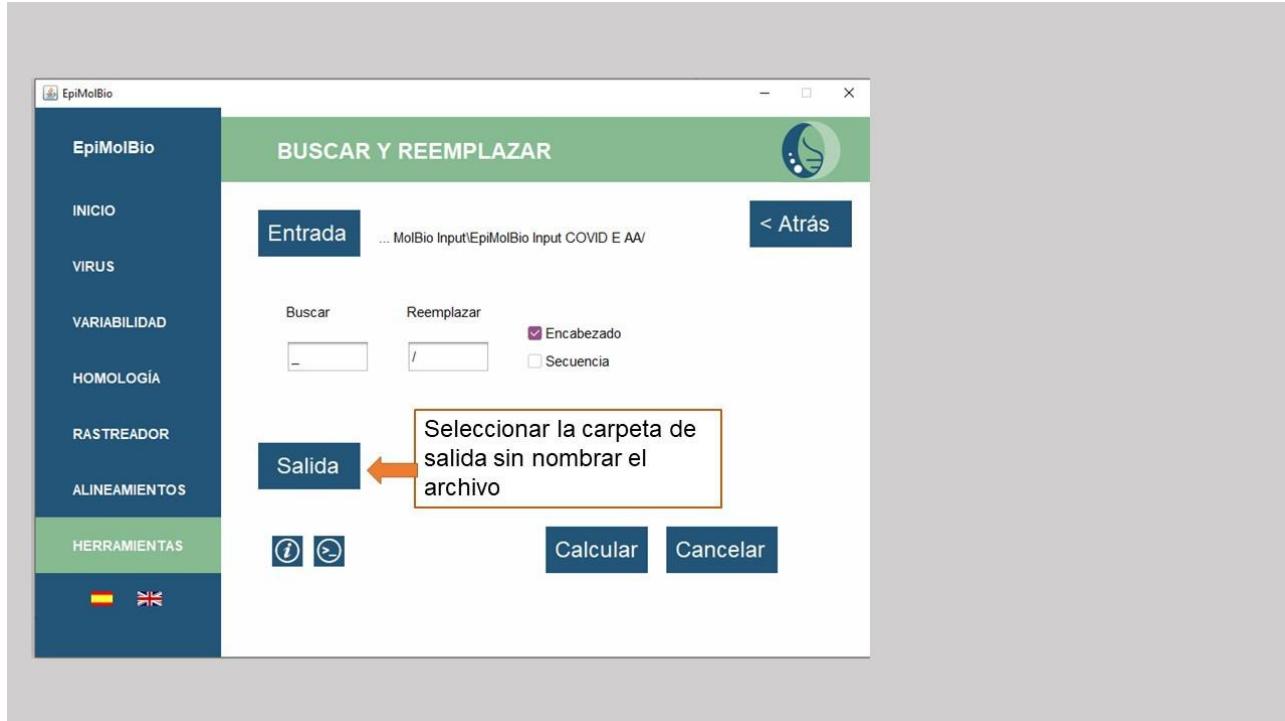
5)



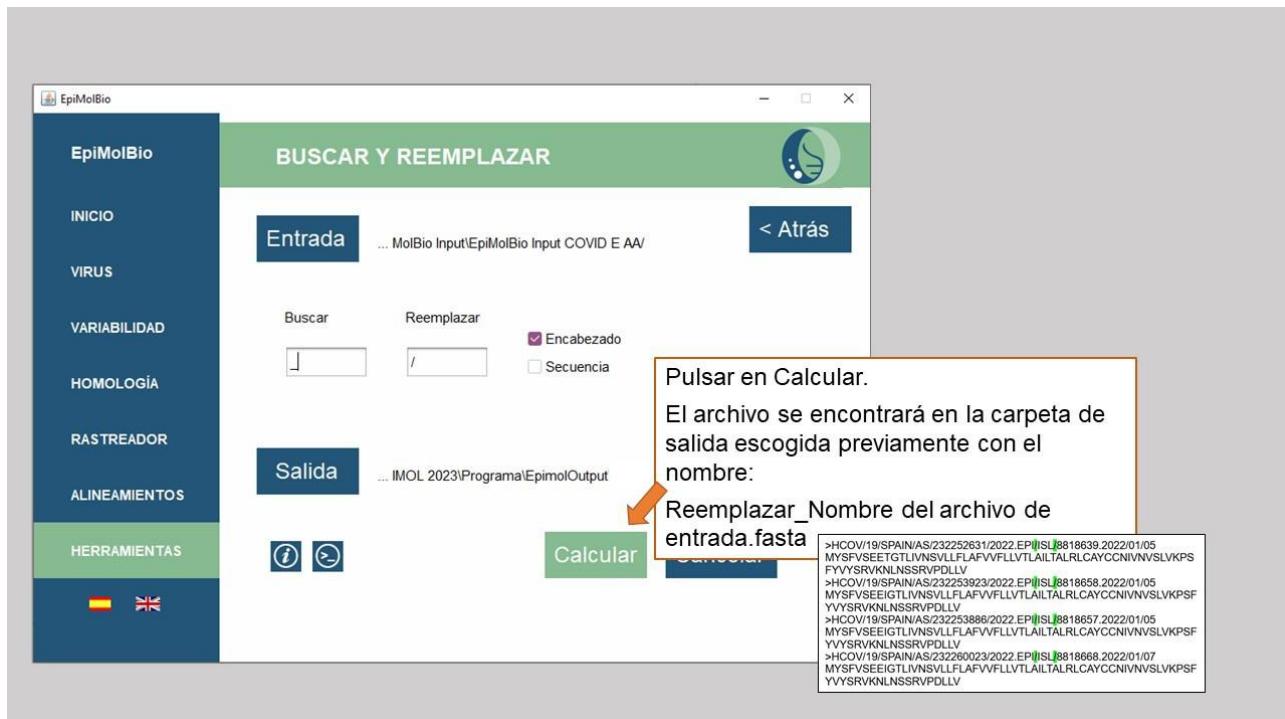
6)



7)



8)



VI.2.FILTROS

VI.2.A) FILTRADO POR ENCABEZADO

Esta herramienta **sirve para filtrar uno o varios archivos en formato “.fasta” usando los parámetros de su encabezado**, separándolos en **archivos distintos según el parámetro** escogido. También permite **eliminar gaps y traducir a la vez**.

Por ejemplo: separar secuencias por su variante, país de origen, año de toma de muestra, nombre o número de acceso a partir del siguiente encabezado: 10_CD.ES.1998.AF000454.IC2258.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiera filtrar. Todos deben el mismo tipo de información en cada parámetro del encabezado y en el mismo orden.

En el campo “**Encabezado**” elegir el ítem por el cual se va a realizar el filtrado. EpiMolBio contempla hasta 5 ítems. En el ejemplo anterior, si se quiere filtrar por año, habría que escoger el ítem 3.

En el campo “**Separador**” introducir el carácter que sirve de separador en el encabezado de las secuencias. En el ejemplo anterior, sería un punto “.”, en otro caso puede ser otro carácter como “_” o “/”.

En el campo “**Opciones**” se puede seleccionar “**Eliminar Gaps**” para eliminar todos los gaps y “**Traducir**” para traducir de nucleótidos a aminoácidos.

En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo sin nombrarlo. Se genera un archivo por cada ítem filtrado que contiene las secuencias .fasta que presentan los mismos caracteres para ese ítem. Éstos se nombran de forma automática de la siguiente forma: “Filtro_Encabezado_Nombre del archivo de entrada.fasta”.

Ejemplo de archivo .fasta de entrada con recombinante 10_CD del VIH-1 para filtrar por año:

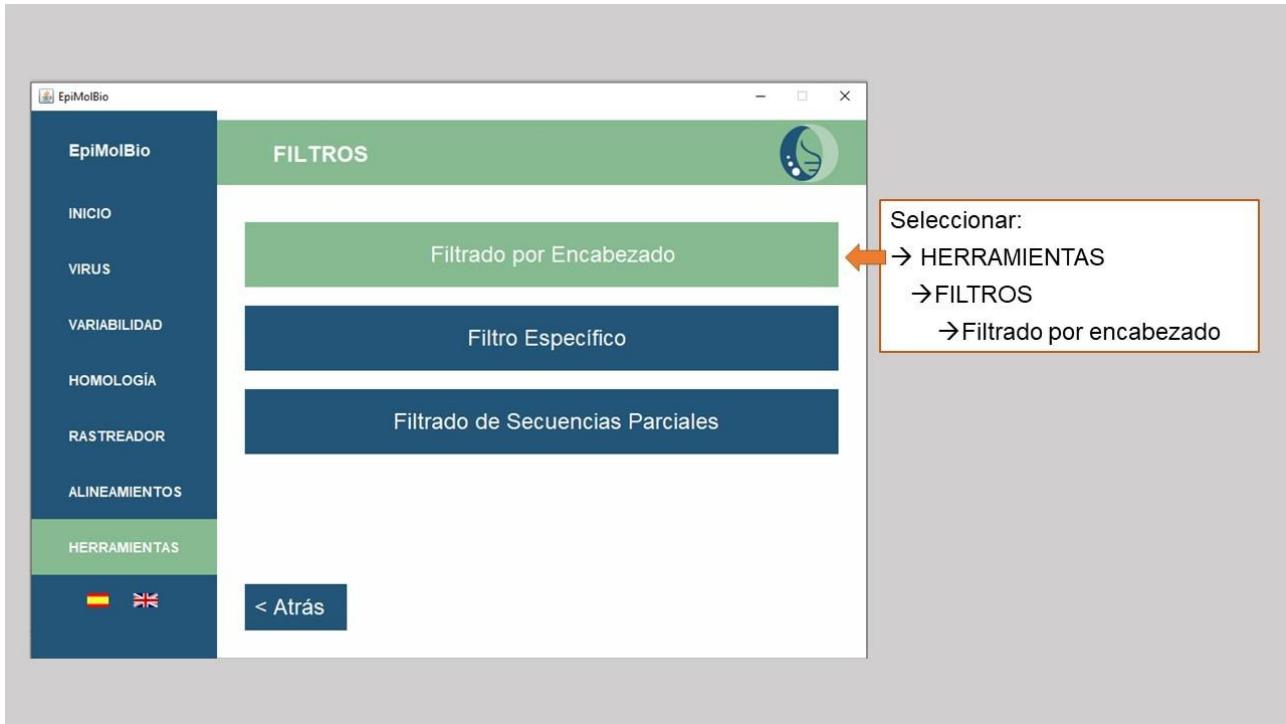
```
>10_CD.TZ.1996 6950.AY036334
PQITLWQRPLTVKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGGFIKVRQYE
QVLIIECGKKAI GTVVGPTPVNIIGRNMLTQIGCTLN
>10_CD.ES.2006 06SP110_320882.EU255456
PQITLWQRPLTIKIGGQLKEAL?DTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYEQIL
IEICGKKAI GTVVGPTPVNIIGRNMLTQIGCTLN
>10_CD.FR.2007_22_csf.FJ549988
PQITLWQRPLVSIKVGGQLKEALLDTGADDTVLEEIKLPGNWPKMIGGIGGFIKVRQYDQI
LIEICGKRAIGTVVGPTPINIIGRNMLTQLGCTLN
>10_CD.TZ.2009_TZ_10_003316_CRF10_CD.HM572362
PQITLWQRPLTVKVGQLKEALLDTGADDTVLEEMN?PGKWKPKMIGGIGGFIKVRQYD
QILVEICGHKAIGTVVGPTPVNIIGRNLLTQIGCTLN
>10_CD.TZ.2009_TZ_09_032645_CRF_10CD.HM572363
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVVEEMCLPGKWKPKMIGGIG?FIKVRQYDQI
```

Ejemplo de archivos de salida con los .fasta agrupados por año:

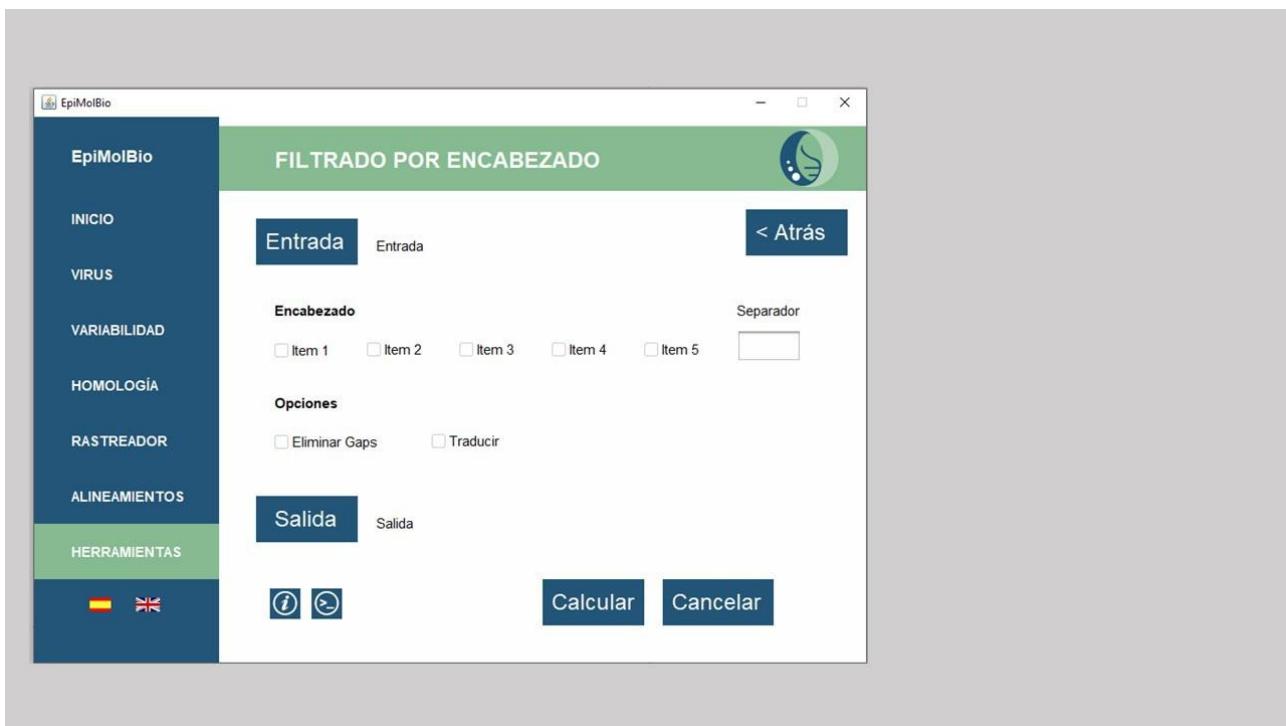
| Nombre | Fecha de modificación | Tipo |
|---|-----------------------|------------|
| Filtro_Encabezado_PR NAIVE_procesado_10_CD_1996 | 22/05/2023 14:21 | FASTA File |
| Filtro_Encabezado_PR NAIVE_procesado_10_CD_2006 | 22/05/2023 14:21 | FASTA File |
| Filtro_Encabezado_PR NAIVE_procesado_10_CD_2007 | 22/05/2023 14:21 | FASTA File |
| Filtro_Encabezado_PR NAIVE_procesado_10_CD_2009 | 22/05/2023 14:21 | FASTA File |

Paso a paso:

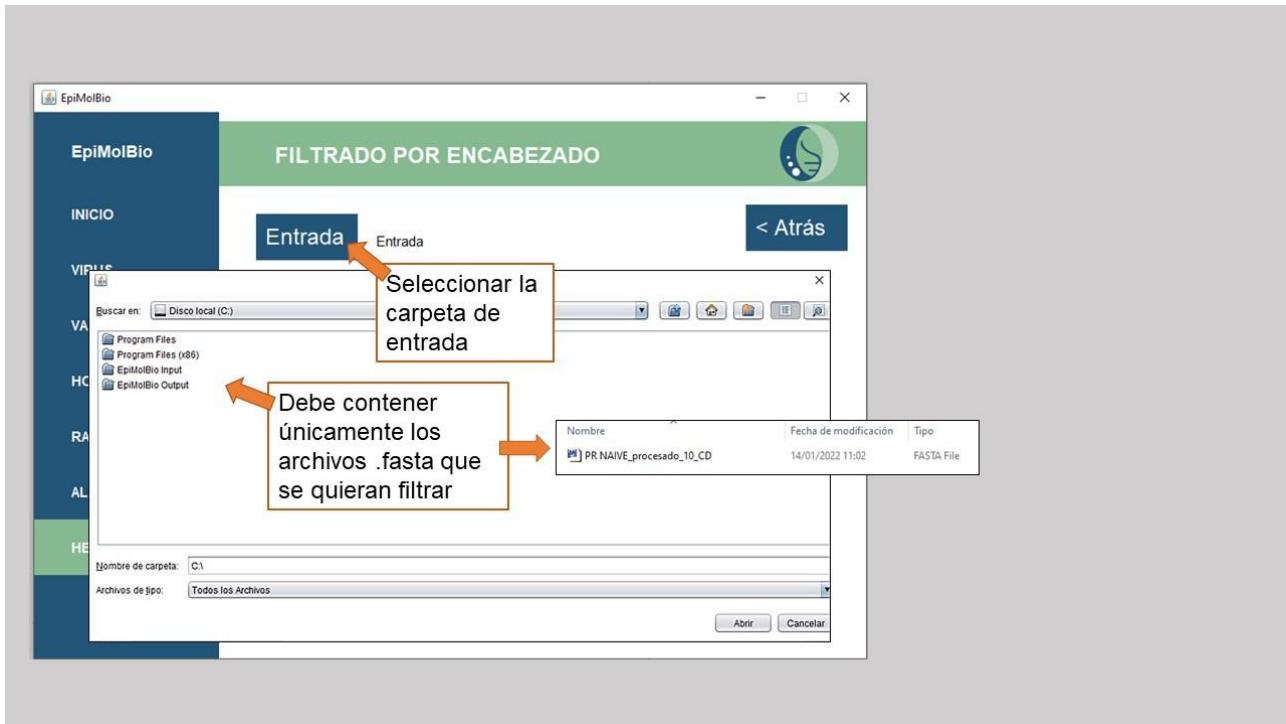
1)



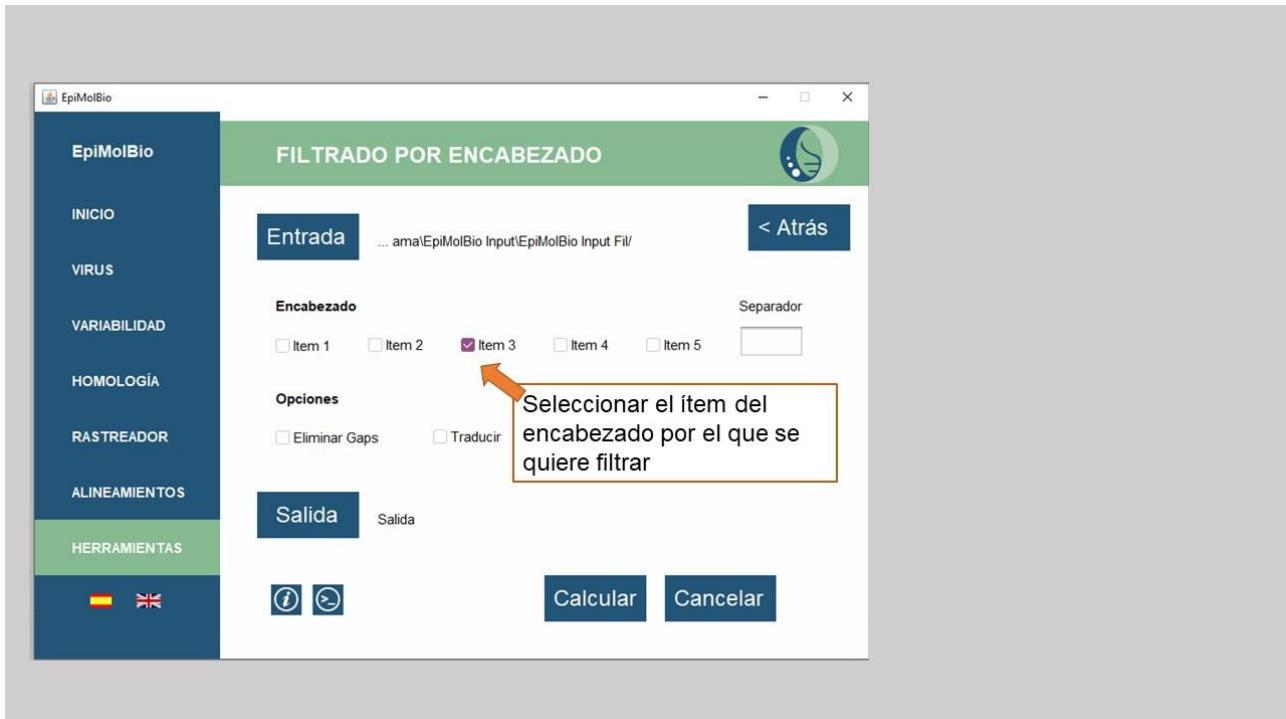
2)



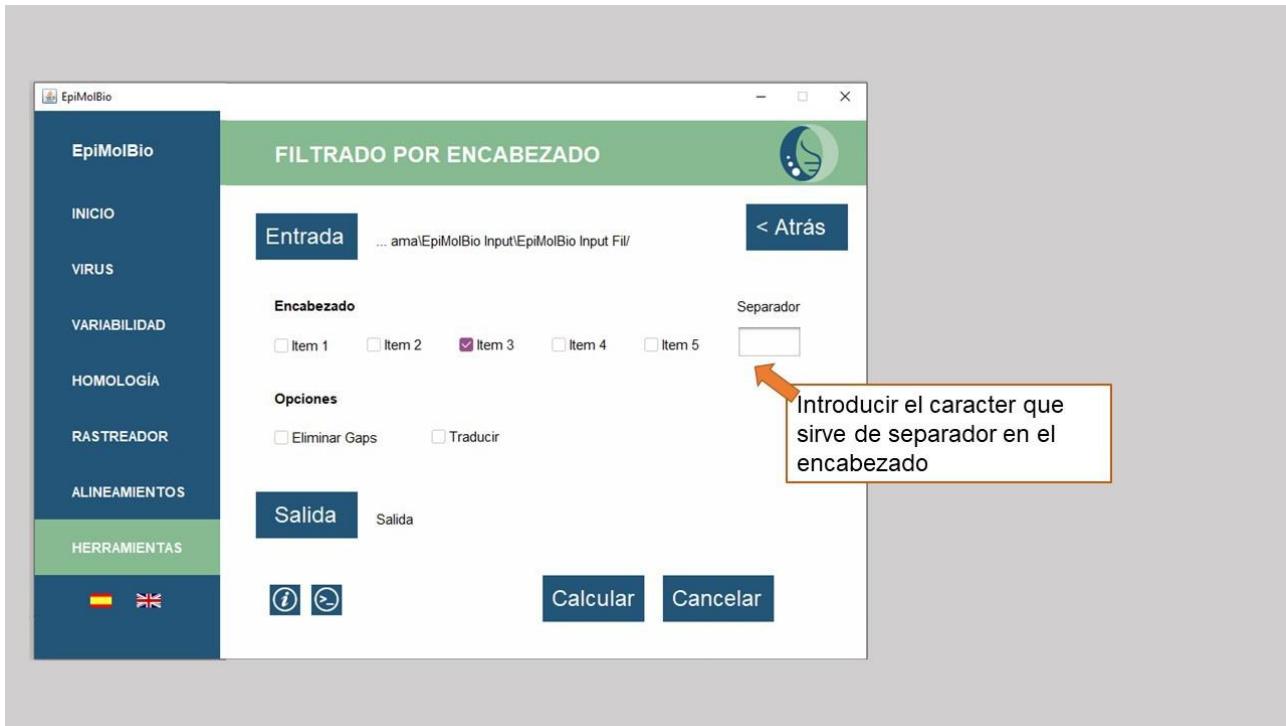
3)



4)



5)



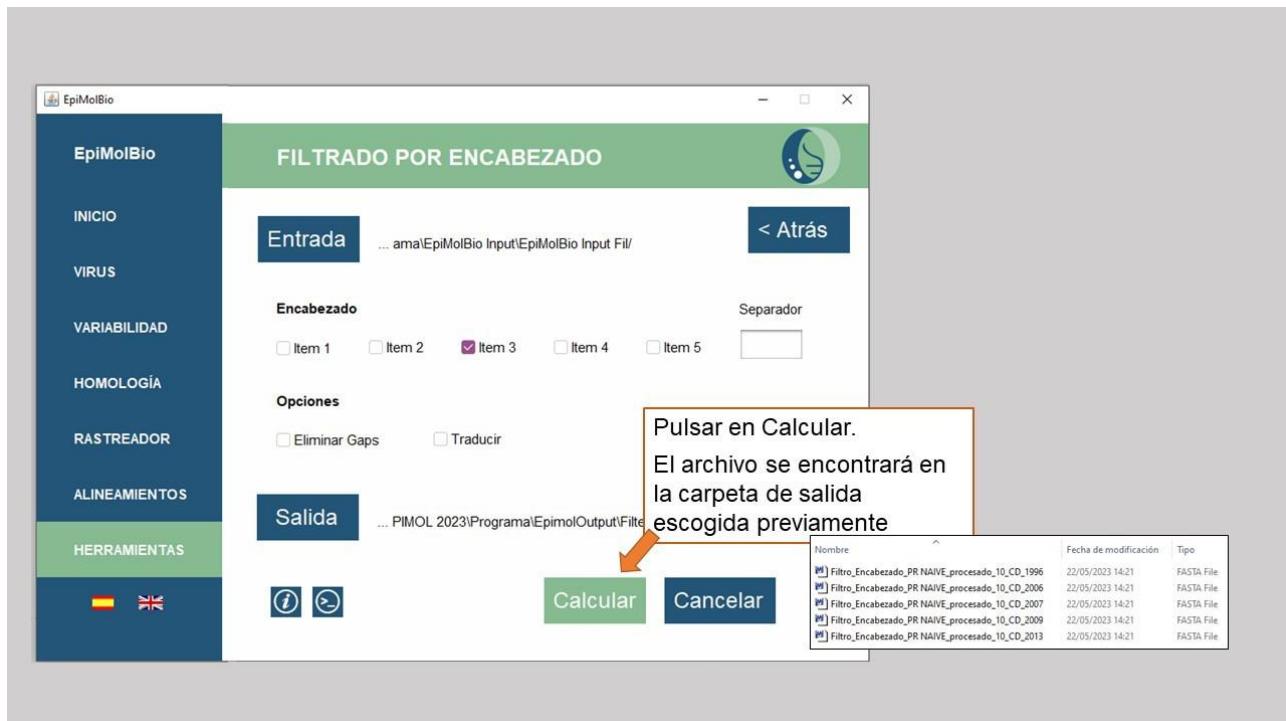
6)



7)



8)



VI.2.B) FILTRO ESPECÍFICO

Esta herramienta sirve para **filtrar secuencias de archivos en formato .fasta que tienan una serie de caracteres concreta en su encabezado.**

Por ejemplo: filtrar secuencias por su variante, país de origen, año de toma de muestra, nombre o número de acceso a partir del siguiente encabezado: 10_CD.TZ.1996.AF000454.IC2258.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiera filtrar. Todos deben el mismo tipo de información en cada parámetro del encabezado y en el mismo orden.

En el campo “**Secuencia de Filtrado**” introducir la serie de caracteres por los que se quiere filtrar junto con el separador previo y posterior. En el ejemplo anterior, si se quiere filtrar por país de origen, eligiendo Tanzania (TZ), introducir “.TZ.”

En **salida** habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo sin nombrarlo. Por cada archivo de entrada se genera un archivo de salida con las secuencias que contengan la serie de caracteres especificada. Los archivos se nombran de forma automática de la siguiente forma: “Filtro_Específico_Nombre del archivo de entrada.fasta”.

Ejemplo de archivo .fasta de entrada con recombinante CD_10 del VIH-1 para filtrar por país de origen (Tanzania o TZ):

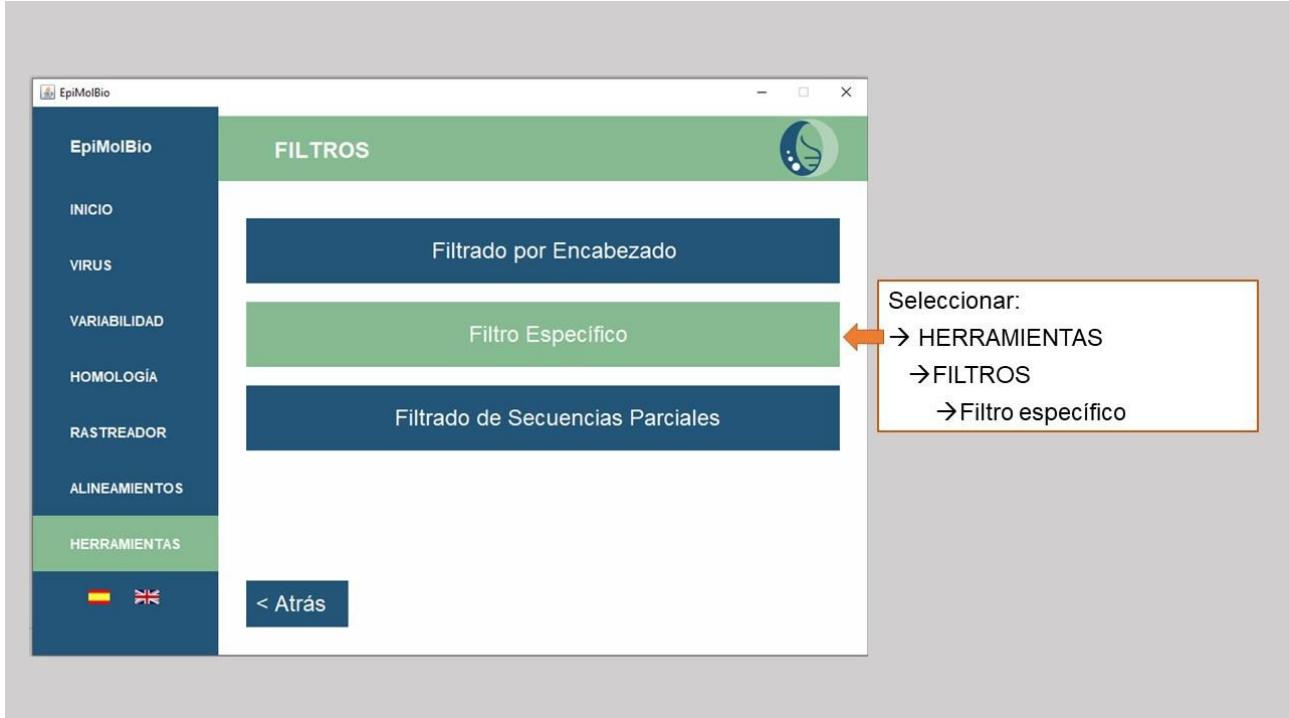
```
>10_CD.TZ.1996.6950.AY036334
PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGG
FIKVROYEQVNLIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>10_CD.ES.2006.06SP110_320882.EU255456
PQITLWQRPLVTIKIGGQLKEAL?DTGADDTVLEEINLPGKWKPKMIGGIGGF
KVRQYEQILIEICGKKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
>10_CD.FR.2007.22_csf.FJ549988
PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEIKLPGNWPKPMIGGIGF
IKVRQYDQILIEICGKRAIGTVLVGPTPINIIGRNMLTQLGCTLNF
>10_CD.TZ.2009.TZ_10_003316_CRF10_CD.HM572362
PQITLWQRPLVTVKVGGQLKEALLDTGADDTVLEEMN?PGKWKPKMIGGIG
GFIKVRQYDQILVEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
>10_CD.TZ.2009.TZ_09_032645_CRF_10CD.HM572363
PQITLWQRPLTIKGQLKEALLDTGADDTVVEEMCLPGKWKPKMIGGIG?
FIKVRQYDQILVEICGHEAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
>10_CD.TZ.2007.TZ_08_017196_CRF10_CD.HM572364
PQITLWQR?LVTVKIEGQLKE?LLDTGADDTVLEDINLPGKWP?MIGGIGG?I
KVRQYDQI?VDICG??A?GTVLVGPTPVNIIGR?LLTQIGCTLNF
>10_CD.TZ.2013.BL_4015.KX775305
PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGG
FIKVRQYDHILIEICGKKTGTVLIGPTPVNIIGRNLLTQIGCTLNF
```

Ejemplo de archivo de salida con los .fasta que contienen “TZ” en su encabezado:

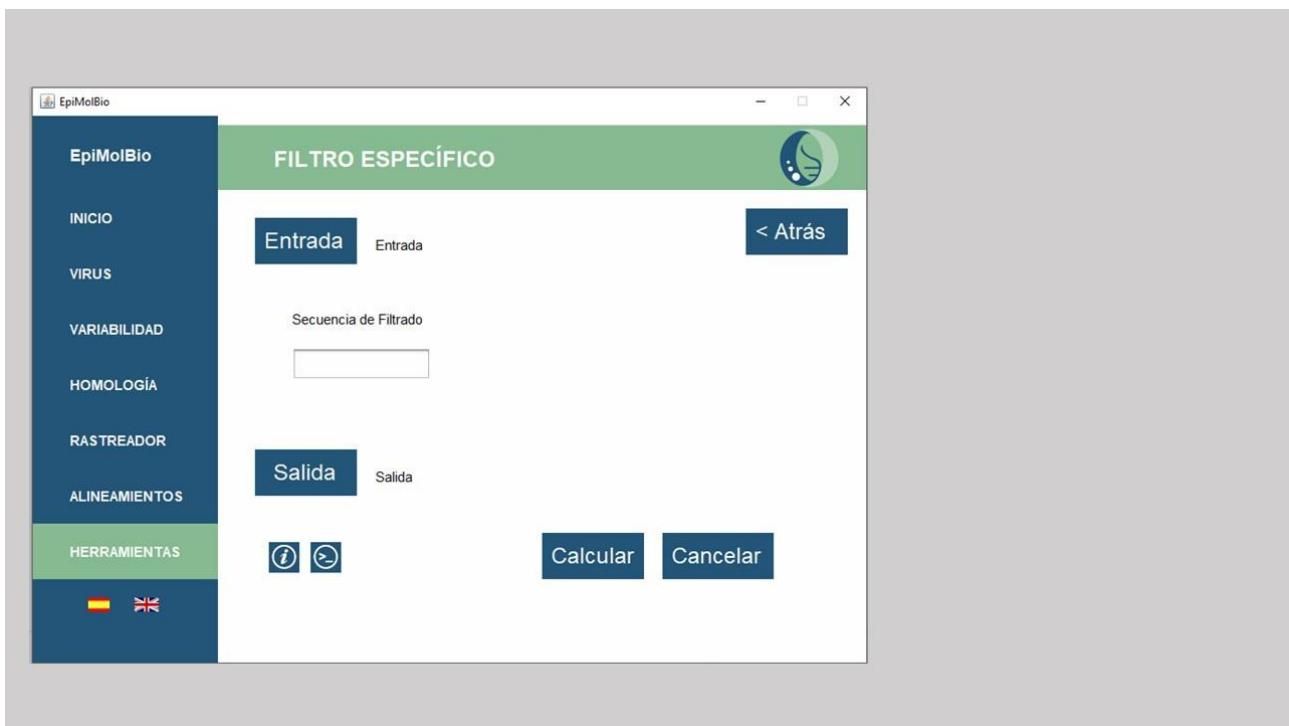
```
>10_CD.TZ.1996.8950.AY036334
PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGG
FIKVROYEQVNLIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>10_CD.TZ.2009.TZ_10_003316_CRF10_CD.HM572362
PQITLWQRPLVTVKVGGQLKEALLDTGADDTVLEEMN?PGKWKPKMIGGIG
GFIKVRQYDQILVEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
>10_CD.TZ.2009.TZ_09_032645_CRF_10CD.HM572363
PQITLWQRPLTIKGQLKEALLDTGADDTVVEEMCLPGKWKPKMIGGIG?
FIKVRQYDQILVEICGHEAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
>10_CD.TZ.2007.TZ_08_017196_CRF10_CD.HM572364
PQITLWQR?LVTVKIEGQLKE?LLDTGADDTVLEDINLPGKWP?MIGGIGG?I
KVRQYDQI?VDICG??A?GTVLVGPTPVNIIGR?LLTQIGCTLNF
>10_CD.TZ.2013.BL_4015.KX775305
PQITLWQRPLVTVKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGG
FIKVRQYDHILIEICGKKTGTVLIGPTPVNIIGRNLLTQIGCTLNF
```

Paso a paso:

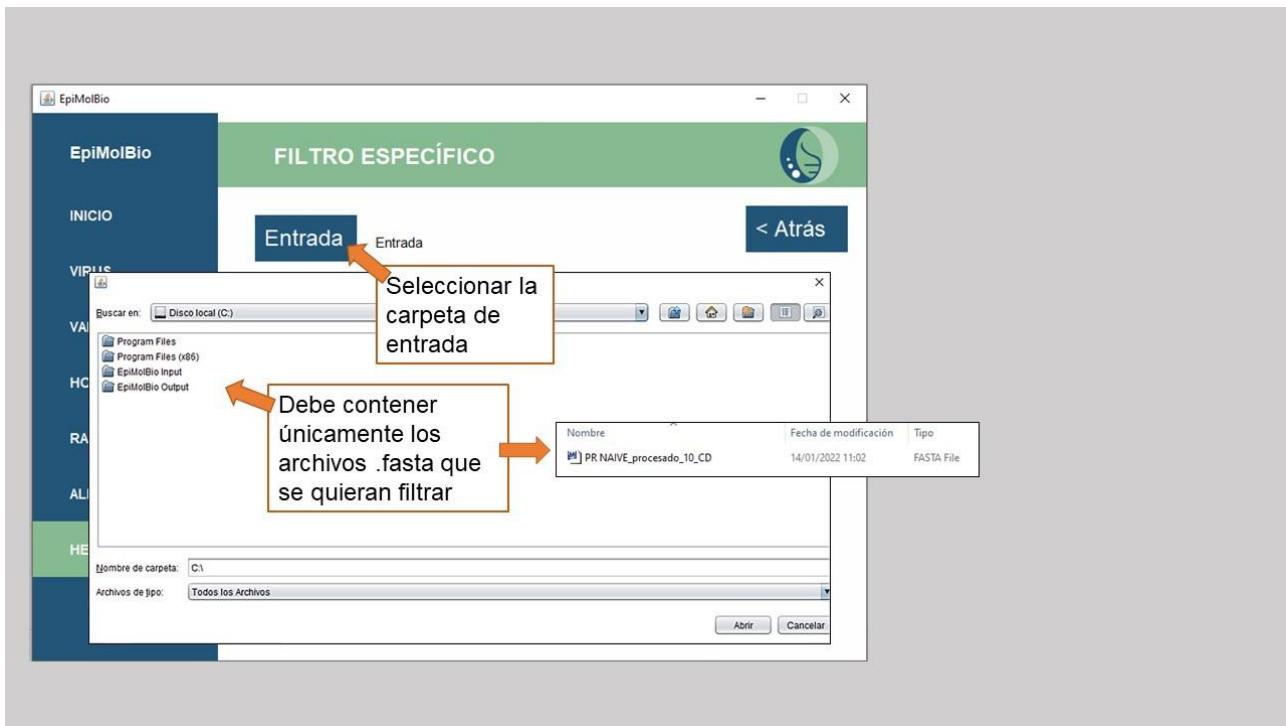
1)



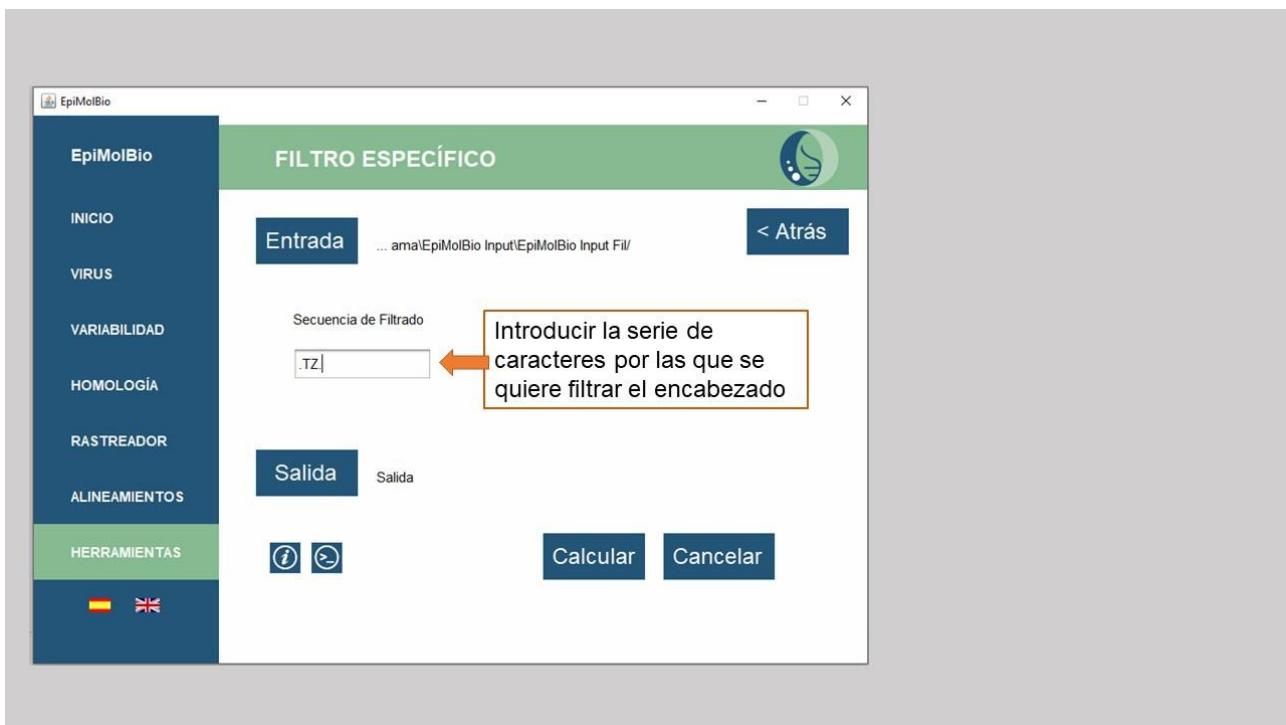
2)



3)



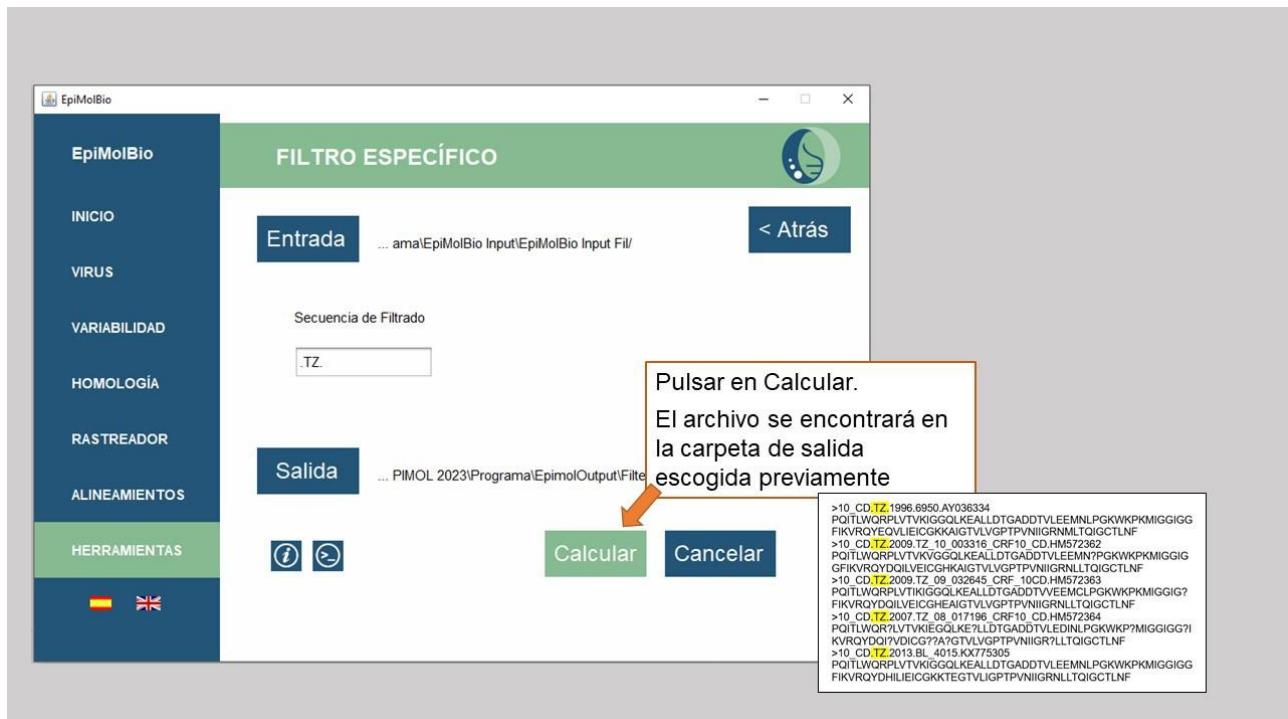
4)



5)



6)



VI.2.C) FILTRADO DE SECUENCIAS PARCIALES

Esta herramienta **permite filtrar secuencias según su calidad**. Esta calidad va en función de la cantidad de residuos desconocidos que contengan las secuencias, apareciendo como “?” en las secuencias en aminoácidos o como “N” en las secuencias en nucleótidos. A mayor cantidad de estos caracteres, menor calidad de la secuencia. La herramienta permite establecer el umbral de calidad para obtener uno o varios archivos .fasta con secuencias que superen dicha calidad. También se obtiene un archivo .html donde se muestran las secuencias que se han perdido al aplicar este filtro.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quieran filtrar.

En el campo “**Tipo de Secuencia**” seleccionar si las secuencias de entrada están en nucleótidos o en aminoácidos.

En el campo “% **Filtrado**” introducir la cifra del porcentaje de filtrado con 1 decimal. Por ejemplo: introduciendo 95.0 en el porcentaje de filtrado, se eliminarán las secuencias que contengan un 5% o más de N del archivo de salida, dejando aquellas que contengan ninguna o <5% de “N”. Introduciendo 100.0, solo quedaran aquellas secuencias sin “N”.

En **salida** se generan dos archivos: un archivo .fasta con las secuencias filtradas y un archivo .html con las secuencias perdidas al aplicar el filtro. Habrá que seleccionar la carpeta de salida donde queremos que aparezca el archivo .fasta sin nombrarlo. Éste se nombrará automáticamente de la siguiente manera: “Filtrado_Parciales_Nombre del archivo de entrada.fasta”. El archivo de salida .html se nombrará automáticamente como “Secuencias_Perdidash.html”.

Ejemplo de un archivo de entrada con 15 secuencias, 7 de ellas con “?“:

```
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPKGKWRPKMIGGIGGFIKVRQYDQILMEICGKKAIGTV
LVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.GA.-MKK27.AM903433
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEEINLPKGKWPKMIGGIGGFIKV?QYDQILIEICGKKAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CI.2001.pg123.AY207737
PQITLWQRPLVTVKIGGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKRAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.ES.2000.SP2756..00.AY248312
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEEINLPKGKWPKMIGGIGGFIKVRQNDQILIEICGKKAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CM.2000.CMNYU3475.AY359728
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKRAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ077.DQ273946
PQITLWQRPLVTVKIGGGQLIEALLDTGADDTVLEEINLPKGKWPKMIGGIGGFIKVRQYDQILIEICGKKAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ085.DQ273952
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQILMEICGKKAIGTVL
VLVGPPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ262.DQ273962
PQITLWQRPLVTVRIGEQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQILIEICGKKAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ270.DQ273962
PQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIIEICGKKAIGTVL
VGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297189
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297190
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297191
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297192
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297193
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
>06_cpx.CM.2001.M4066N.DQ297194
PQITLWQRPLVTARIGGQLIEALLDTGADDTVLEDINLPKGKWPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVL
VGPTP?NIIGRNMLTQIGCTLNF
```

Ejemplo del archivo de salida .fasta aplicando el filtro 100%, donde se muestran las 8 secuencias sin ningún “?”:

```
>06_cpx.ES.2000.19804.AF354004
PQITLWQRPLTVRIGGQLIEALLDTGADDTVLEDINLPGKWRPKMIGGIGGFIVKRQYDQI
LMEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CI.2001.pc123.AY207737
PQITLWQRPLTVKIGGQLIEALLDTGADDTVLEDINLPGKWPKMIGGIGGFIVKRQYDQI
PIEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.ES.2000.SP2756_00.AY248312
PQITLWQRPLTVRIGGQLIEALLDTGADDTVLEEINLPGKWPKMIGGIGGFIVKRQNDQI
LIEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.CM.2000.00CMNYU3475.AY359728
PQITLWQRPLTVRIGGQLIEALLDTGADDTVLEDINLPGKWPKMIGGIGGFIVKRQYDQI
PIEICGKRAIGTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ077.DQ273946
PQITLWQRPLTVKIGGQLIEALLDTGADDTVLEEINLPGKWPKMIGGIGGFIVKRQYDQI
IEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ085.DQ273952
PQITLWQRPLTVKVGQQLEALLDTGADDTVLEDINLPGKWPKMIGGIGGFIVKRQYDQI
LMEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ262.DQ273960
PQITLWQRPLTVRIGEQLIEALLDTGADDTVLEDINLPGKWPKMIGGIGGFIVKRQYDQI
IEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
>06_cpx.NG.-DDJ270.DQ273962
PQITLWQRPLTVRIGGQLIEALLDTGADDTVLEDINLPGKWPKMIGGIGGFIVKRQYDQI
HIEICGKKAI GTVLVGPTPVNIIGRNMLTQIGCTLNF
```

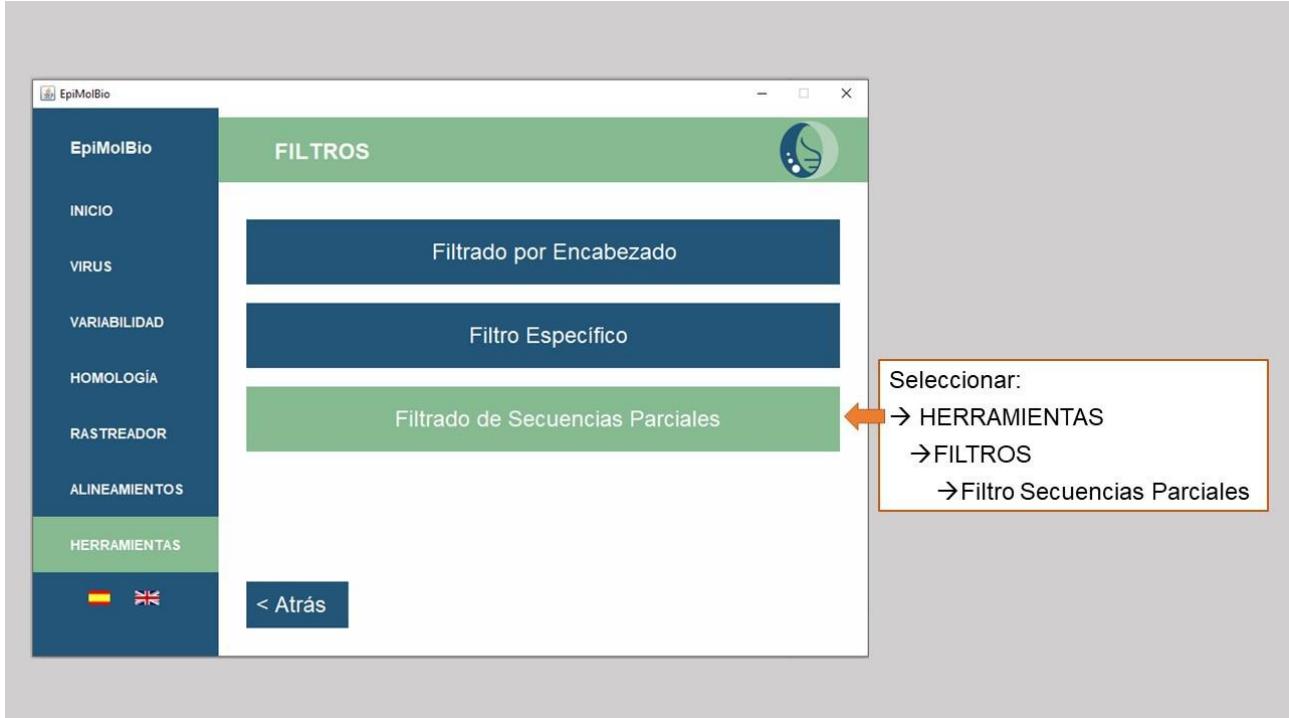
Ejemplo del archivo de salida .html aplicando el filtro 100% donde se muestra el número y porcentaje de las secuencias perdidas (las 7 secuencias con “?”):

| Filtrado de Secuencias Parciales | | | | |
|---|--------------------|------------------------|---------------------|---------------------|
| Archivo | Secuencias Totales | Secuencias Recuperadas | Secuencias Perdidas | Porcentaje Perdidas |
| 06_cpx.fasta | 15 | 8 | 7 | 46.666% |
| Total | 15 | 8 | 7 | 46.666% |

En el archivo de salida .html aparece, en la parte superior, el título del análisis. En la columna “Archivo”, aparece el nombre del archivo de entrada; “Secuencias Totales” muestra el número total de secuencias de entrada; la columna “Secuencias Recuperadas” muestra el número de secuencias en el archivo de salida; “Secuencias Perdidas” muestra el número de secuencias eliminadas por no cumplir los criterios de calidad y la columna “Porcentaje Perdidas” muestra el porcentaje de secuencias perdidas con respecto al total coloreado según el código de colores descrito en Generalidades, que puede consultarse en el archivo de salida .html pulsando en el símbolo azul.

Paso a paso:

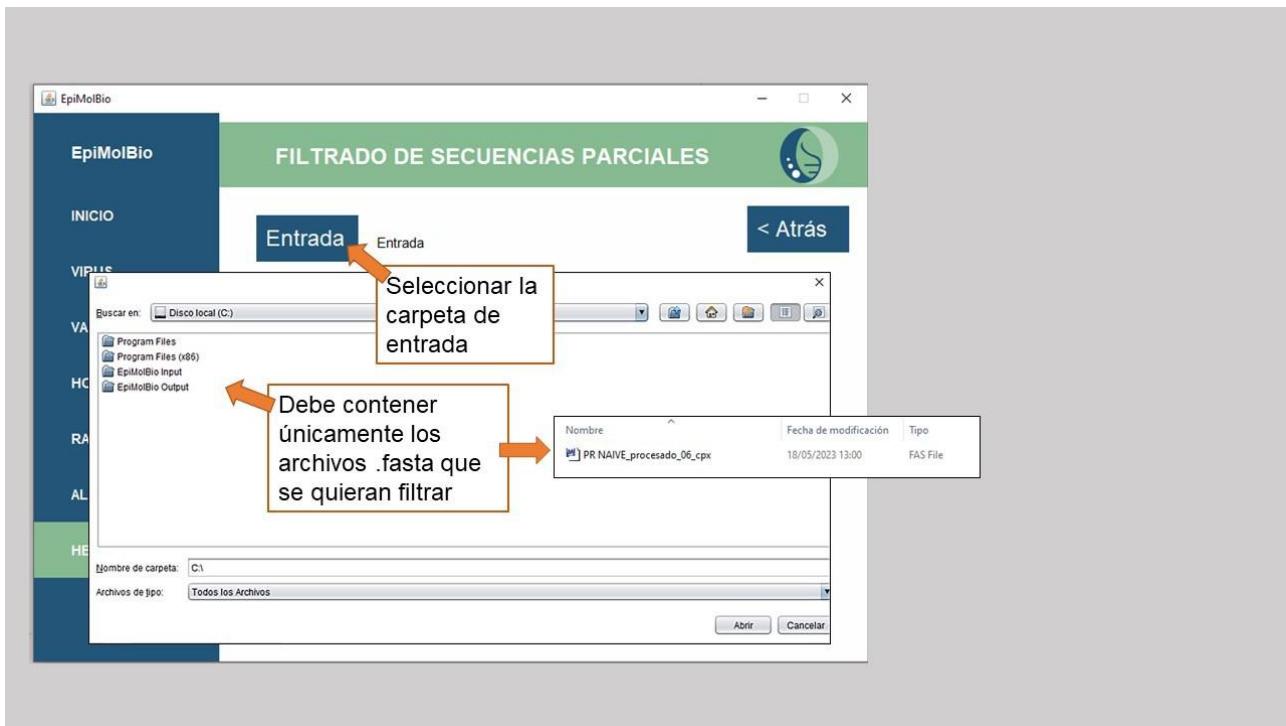
1)



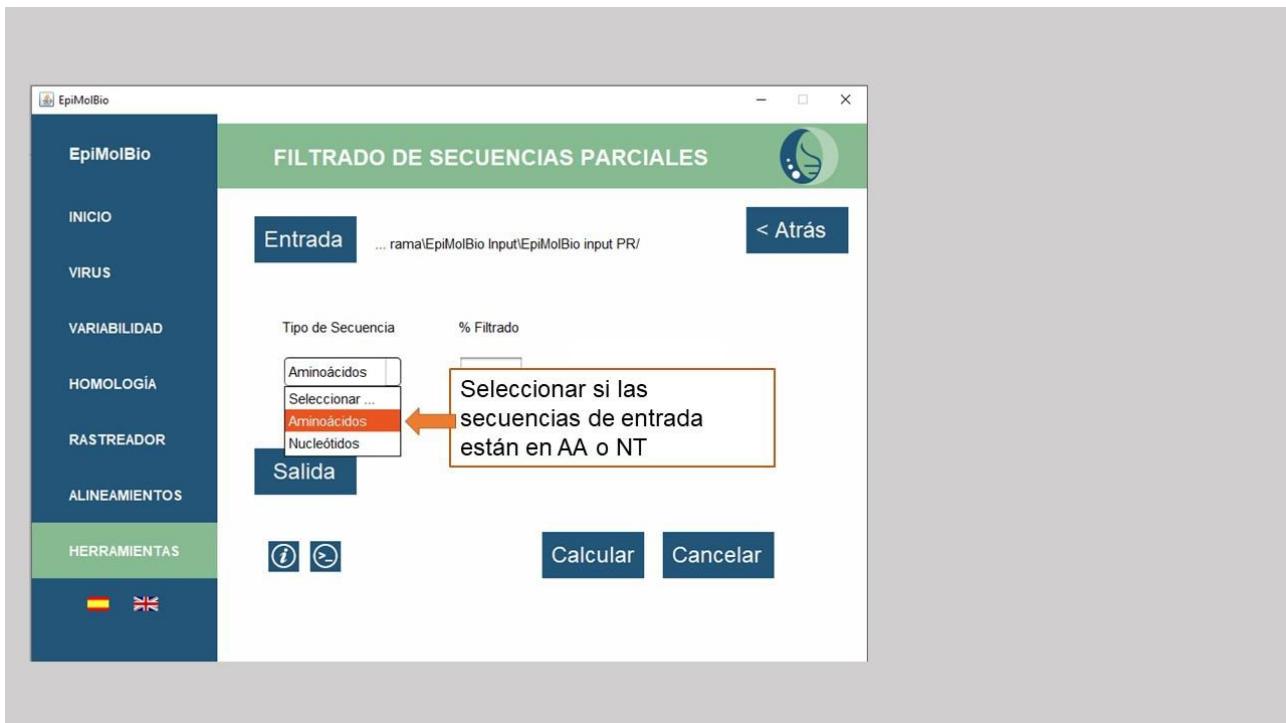
2)



3)



4)



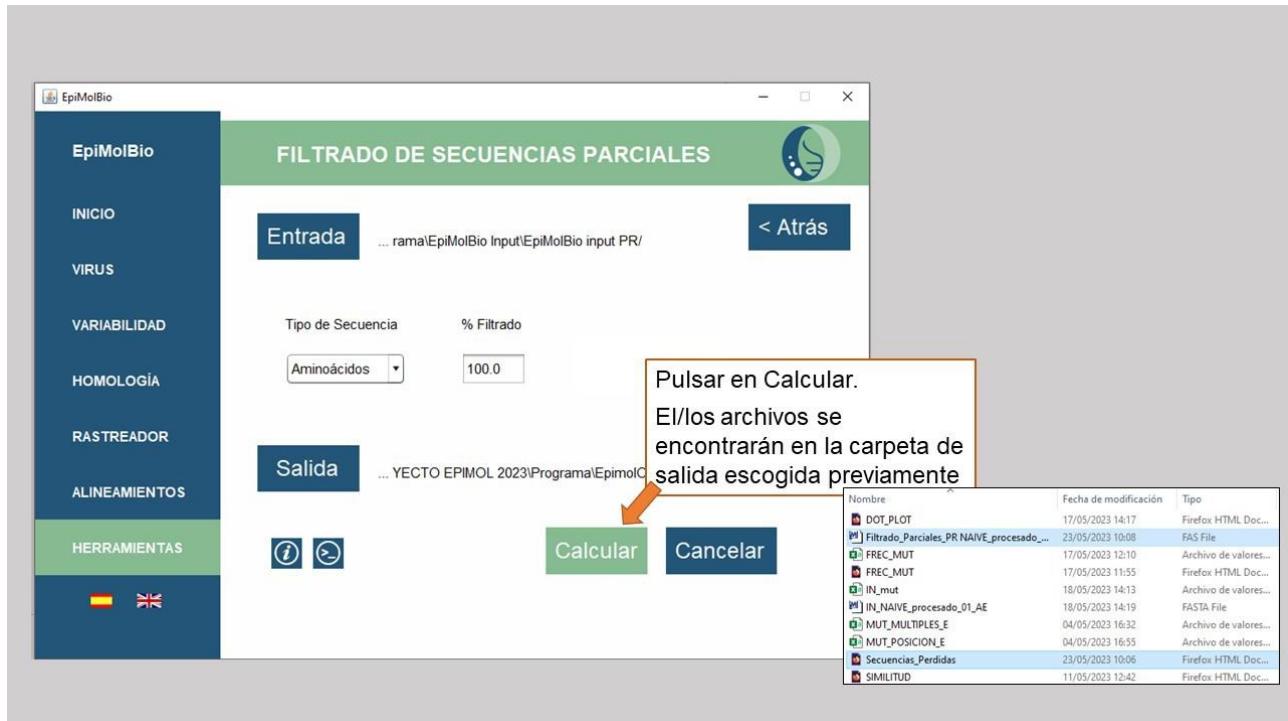
5)



6)



7)



VI.3.TRADUCCIÓN

Esta herramienta **permite traducir secuencias de nucleótidos de archivos .fasta a aminoácidos**. Además permite eliminar los gaps de las secuencias.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta en nucleótidos que se quieran traducir.

Seleccionar la caja “**Traducir**” para llevar a cabo la traducción.

Se puede seleccionar la caja “**Eliminar gaps**” para eliminar todos los gaps de forma automática.

En el campo “**Marco**” escoger el marco de lectura entre marco 1, 2 o 3 para establecer el primer nucleótido donde se comienza a contar codones. En general se emplea el Marco 1, si se asume que las secuencias empleadas están en fase de lectura correcta.

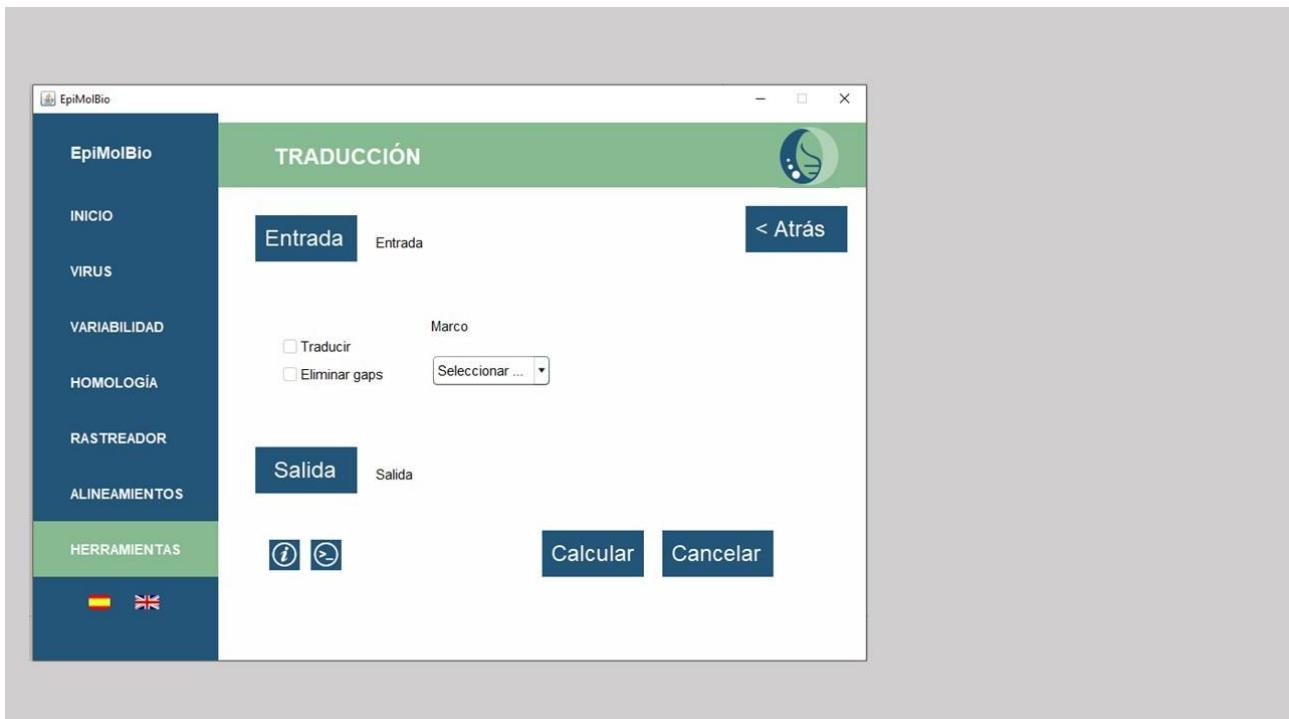
El archivo de salida será un archivo .fasta con las secuencias traducidas a aminoácidos. En **salida** habrá que seleccionar la carpeta de salida donde se quiere que aparezca el archivo .fasta sin nombrarlo. Este archivo se nombra automáticamente de la siguiente manera: “Traducido_Nombre del archivo de entrada.fasta” o “Traducido_Sin_Gaps_NOMBRE del archivo de entrada.fasta” si se ha escogido la opción de eliminar gaps.

Paso a paso:

1)



2)



3)



4)



5)



6)



7)



VI.4. CONTAR SECUENCIAS

Esta herramienta sirve para **contar en número total de secuencias en uno o varios archivos .fasta** o bien **cuántas de esas secuencias contienen mutaciones** con respecto a una secuencia de referencia.

El archivo de **entrada** debe ser la carpeta que contenga exclusivamente los archivos .fasta que se quiera contar.

En el campo “**Formato**” se puede escoger entre las dos funciones. La opción “**Tabla**”, genera una tabla .csv que cuenta todas las secuencias del archivo de entrada. La opción “**Secuencias Mutadas**”, genera una tabla .csv que cuenta sólo las secuencias que presentan mutaciones con respecto a la secuencia de referencia introducida. Si se escoge “Tabla”, el siguiente paso será establecer la carpeta de salida. Si se escoge “Secuencias Mutadas” habrá que llenar los siguientes campos:

En el campo “**Referencia**” introducir la secuencia de referencia, sin espacios ni saltos de línea.

Escoja entre las casillas **AA** y **NT** según si las secuencias del archivo de entrada están en nucleótidos (NT) o aminoácidos (AA).

En **salida**, seleccionar la carpeta de salida donde queremos que aparezca el archivo .csv y nombrar el archivo escribiendo .csv al final. Los formatos de salida .csv pueden abrirse con Excel.

Ejemplo de formato de salida **Tabla** de la herramienta Contar Secuencias:

| | A | B |
|----|------------------|----------------------|
| 1 | Archivo | Número de Secuencias |
| 2 | PR_01_AE.fasta | 26849 |
| 3 | PR_02_AG.fasta | 9577 |
| 4 | PR_03_A6B.fasta | 310 |
| 5 | PR_04_cpx.fasta | 15 |
| 6 | PR_05_DF.fasta | 24 |
| 7 | PR_06_cpx.fasta | 746 |
| 8 | PR_07_BC.fasta | 10916 |
| 9 | PR_08_BC.fasta | 2348 |
| 10 | PR_09_cpx.fasta | 94 |
| 11 | PR_100_01C.fasta | 5 |
| 12 | PR_101_01B.fasta | 4 |

El formato de salida **Tabla** consiste en una tabla .csv. En la tabla se muestra en la primera columna el nombre de los archivos de entrada y en la segunda, el número total de secuencias por archivo. Al final de la tabla se indica el total de secuencias en todos los archivos.

Ejemplo de formato de salida **Secuencias Mutadas** de la herramienta Contar Secuencias:

| | A | B | C | D |
|----|------------------|---------|----------------------|------------|
| 1 | Archivo | Mutadas | Número de Secuencias | Porcentaje |
| 2 | PR_01_AE.fasta | 26849 | 26849 | 100.00% |
| 3 | PR_02_AG.fasta | 9577 | 9577 | 100.00% |
| 4 | PR_03_A6B.fasta | 310 | 310 | 100.00% |
| 5 | PR_04_cpx.fasta | 15 | 15 | 100.00% |
| 6 | PR_05_DF.fasta | 24 | 24 | 100.00% |
| 7 | PR_06_cpx.fasta | 746 | 746 | 100.00% |
| 8 | PR_07_BC.fasta | 10916 | 10916 | 100.00% |
| 9 | PR_08_BC.fasta | 2348 | 2348 | 100.00% |
| 10 | PR_09_cpx.fasta | 94 | 94 | 100.00% |
| 11 | PR_100_01C.fasta | 5 | 5 | 100.00% |
| 12 | PR_101_01B.fasta | 4 | 4 | 100.00% |

El formato de salida **Secuencia Mutadas** consiste en una tabla .csv. En la tabla se muestra el nombre de los archivos de entrada en la primera columna; el total de secuencias que presentan mutaciones con respecto a la secuencia de referencia en la segunda columna; el total de secuencias por archivo de entrada en la tercera, y la frecuencia de las secuencias mutadas. en la cuarta.

En este ejemplo, al ser secuencias de un virus de ARN (VIH) con alta variabilidad genética y frecuencia de mutación, todas las secuencias contienen, al menos, 1 mutación con respecto a la secuencia de referencia (aislado HXB2 del HIV).

Paso a paso:

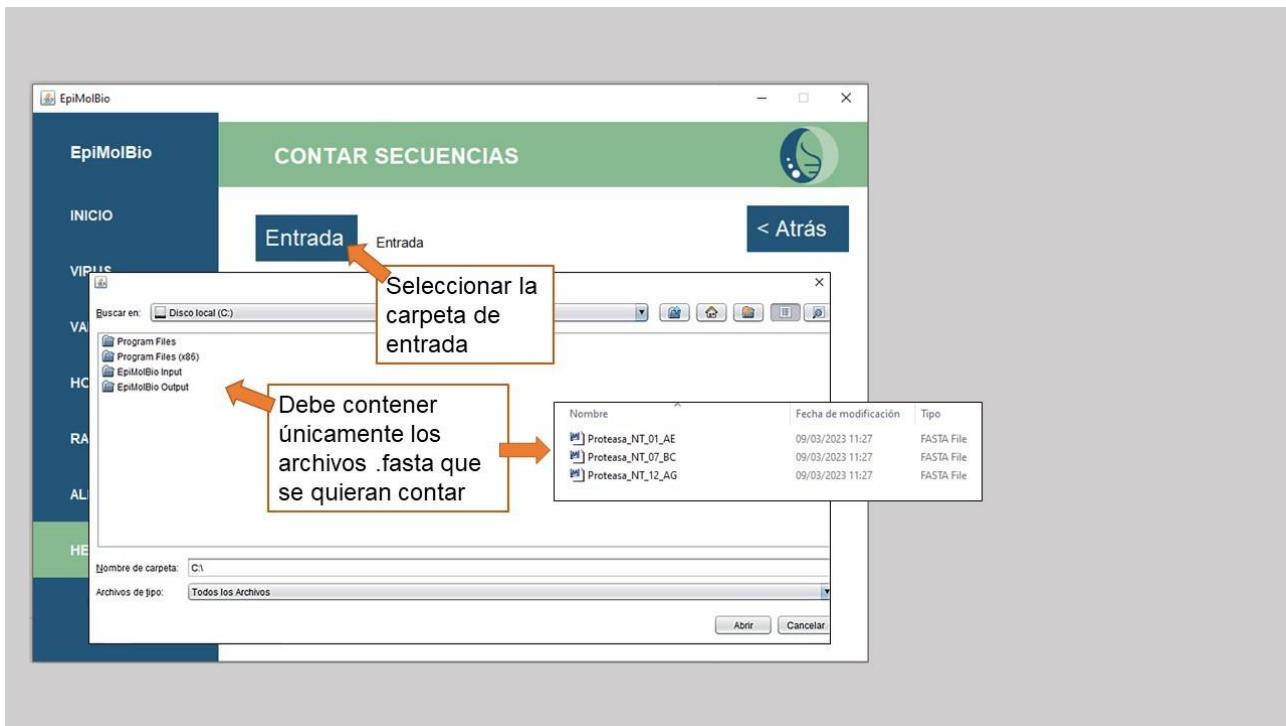
1)



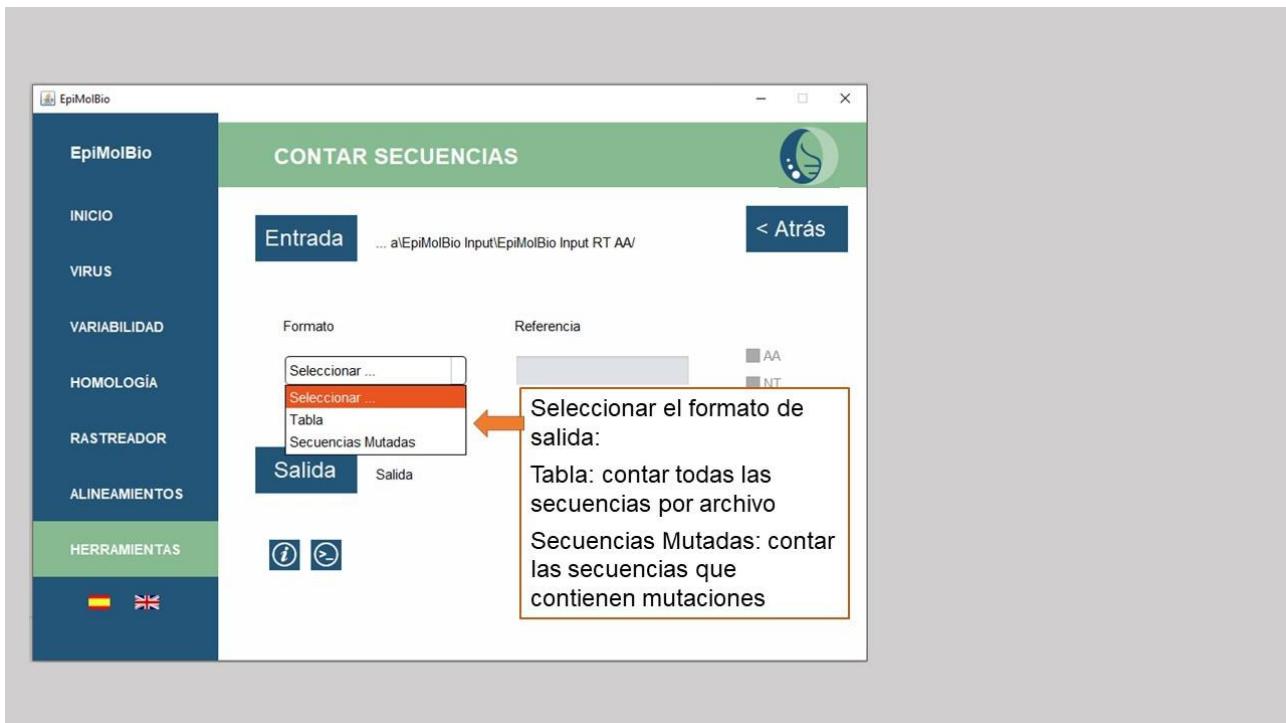
2)



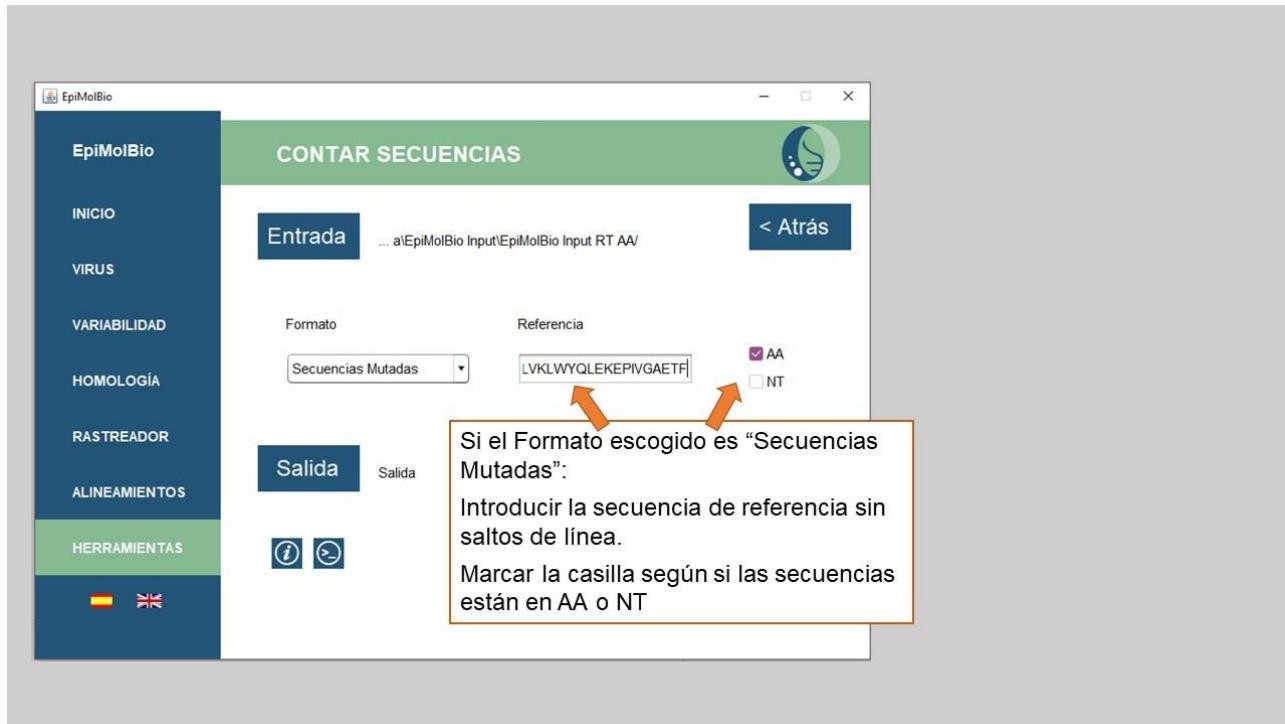
3)



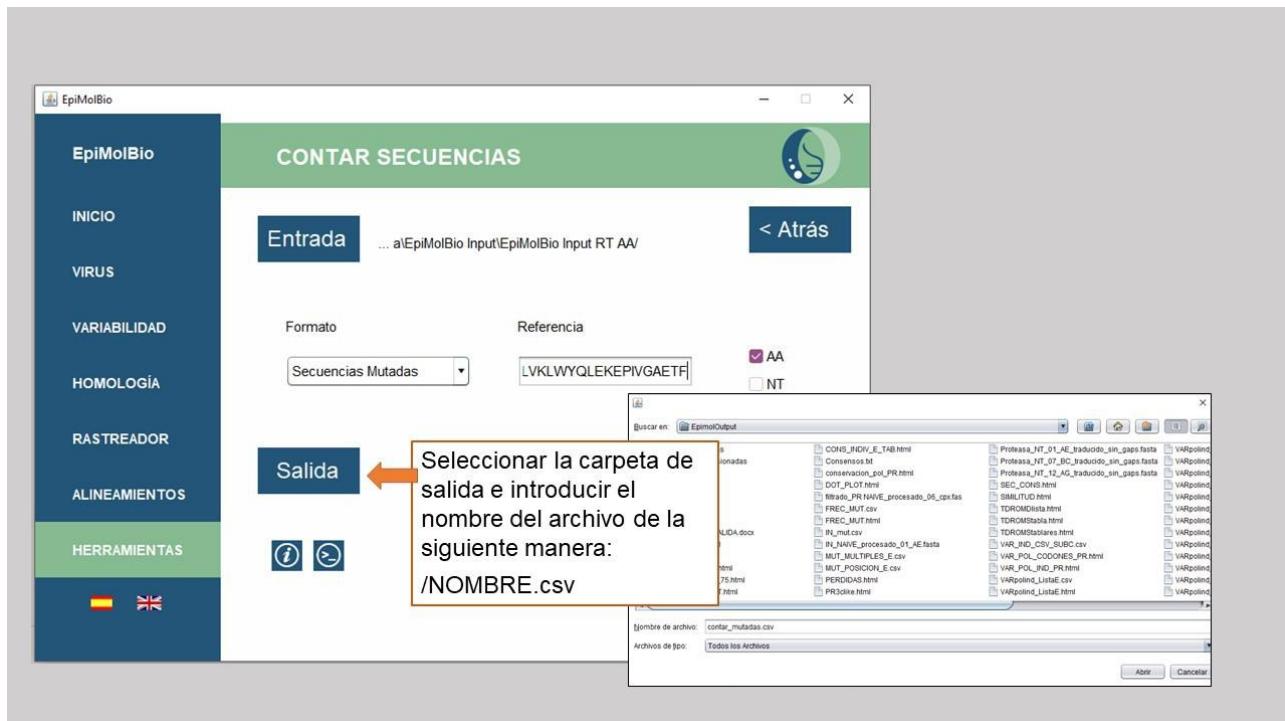
4)



5)



6)



7)



VI.5 PROGRAMAR FUNCIONES

Esta herramienta **permite automatizar las funciones del programa EpiMolBio encadenándolas para que se lleven a cabo sin intervención y de forma secuencial**. Se recomienda su uso cuando sea necesario repetir un mismo proceso o para ejecutar funciones que tarden mucho tiempo o requieran varias funciones, como por ejemplo, procesar un gran volumen de secuencias.

Para llevar a cabo esta tarea es necesario contar previamente con un **archivo .txt** que contenga las **instrucciones** que llevará a cabo de forma automática esta herramienta.

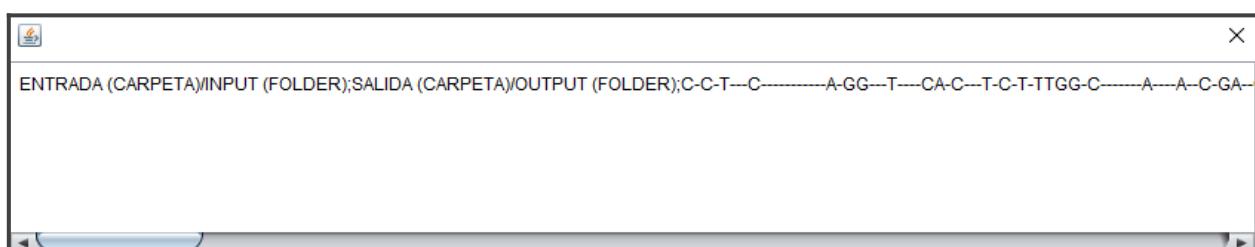
En el siguiente ejemplo automatizaremos las funciones de “Eliminar inserciones” (dentro de Alineamientos) y “Traducción” (en Herramientas) para obtener secuencias traducidas y sin inserciones como salida.

Para crear el primer **archivo de texto**, seleccionar la primera función que queremos que se automaticé sin incluir entrada ni salida.

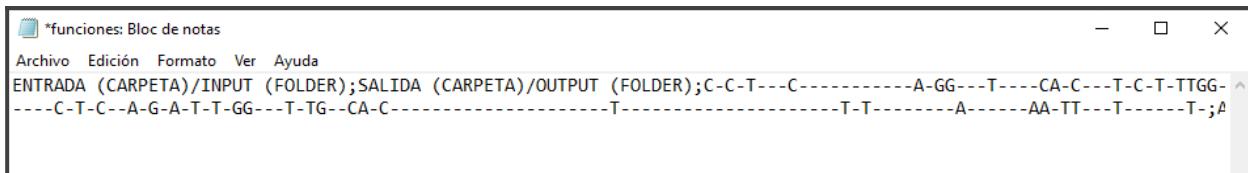
En nuestro ejemplo, en la función “Eliminar Inserciones”, rellenar los parámetros de la función sin incluir entrada ni salida. En este caso, sólo habrá que introducir la secuencia de referencia con gaps, sin saltos de línea ni espacios en AA o NT según los archivos de entrada.



Una vez hecho esto, hacer clic en el botón de programación de funciones. Se abrirá la siguiente ventana:

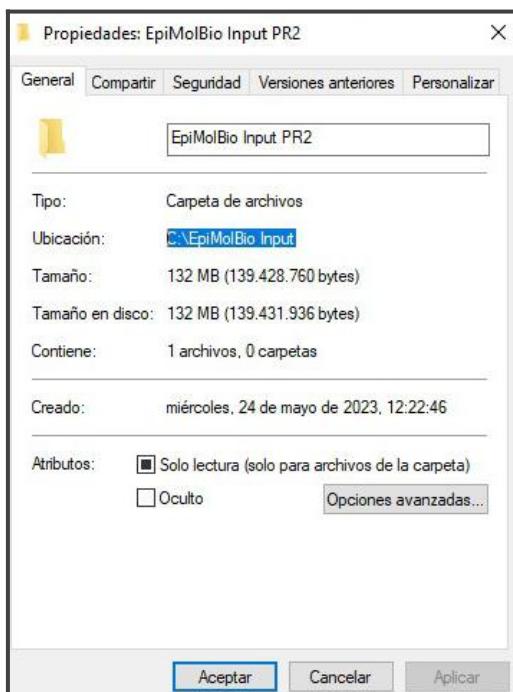


Copiar el contenido del botón y pegarlo en un archivo de texto de la siguiente manera:



```
*funciones: Bloc de notas
Archivo Edición Formato Ver Ayuda
ENTRADA (CARPETA)/INPUT (FOLDER);SALIDA (CARPETA)/OUTPUT (FOLDER);C-C-T---C-----A-GG---T---CA-C---T-C-T-TGGA-----C-T-C--A-G-A-T-T-GG---T-TG---CA-C-----T-----A-----AA-TT---T-----T;A
```

En el archivo de texto generado, sustituir **ENTRADA (CARPETA)/INPUT (FOLDER)** por la ruta de la carpeta que contiene los archivos que se van a analizar. Este puede consultarse en Windows con el botón derecho del ratón, haciendo clic en “Propiedades”.



Hay que copiar tanto lo que está escrito detrás de “Ubicación” como el nombre de la carpeta que aparece arriba en el recuadro, todo ello seguido de “/”

En este ejemplo, quedaría de la siguiente manera: **C:\EpiMolBio Input/EpiMolBio Input PR2;SALIDA...**

Es importante mantener “;” sin borrarlos del archivo de texto.

En Linux se puede copiar directamente el archivo y pegar en el txt, se pegará la ruta del archivo.

Lo siguiente es modificar la salida en el archivo de texto de la misma manera que la entrada, sustituyendo **SALIDA (CARPETA)/OUTPUT (FOLDER)** por la ruta de la carpeta donde se quiera guardar el resultado de la primera función. No es necesario crear la carpeta de antemano, si se **escribe** directamente la ruta, el programa creará la carpeta de forma automática.

Creando la carpeta “Resultado” quedaría de la siguiente manera: **C:\EpiMolBio Input/Resultado;**

Ejemplo del archivo de texto con la entrada y la salida modificadas para la función “Eliminar Inserciones”:

```
C:\EpiMolBio Input/EpiMolBio Input PR2/;C:\EpiMolBio Input/Resultado/;C-C-T---C-----A-GG---T---CA-C---T-C-T-T ^
-A-----C-T-C--A-G-A-T-T-GG---T-TG--CA-C-----T-----T-T-----A-----AA-TT---T-----
```

Una vez automatizada la primera función, se procede a automatizar una segunda. Para ello, la siguiente instrucción debe comenzarse en una nueva línea de texto. Se repetirán los pasos anteriores, sustituyendo la entrada y la salida como se ha explicado previamente, siempre manteniendo “;”.

En nuestro ejemplo, una vez eliminadas las inserciones, se procede a traducir:



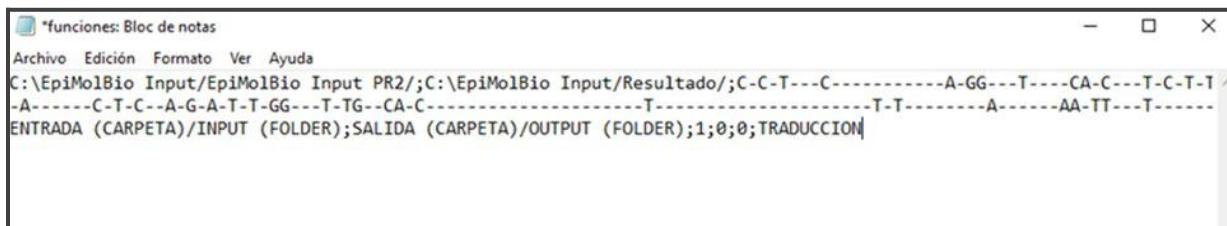
Se completa la información en el interfaz de la función sin introducir la entrada ni la salida. En este caso, se pincha en la caja “Traducir” y se escoge el Marco 1 de lectura.

A continuación, hacer clic en el **botón de programación de funciones** para copiar su contenido:



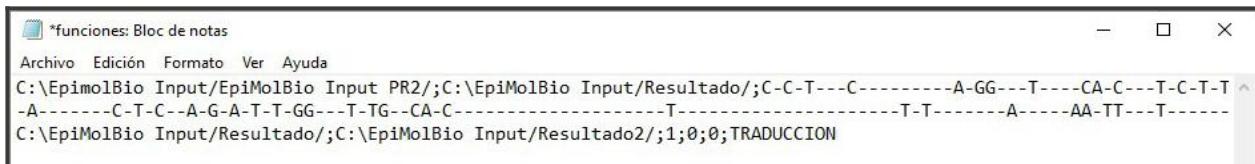
```
ENTRADA (CARPETA)/INPUT (FOLDER);SALIDA (CARPETA)/OUTPUT (FOLDER);1;0;0;TRADUCCION
```

Pegar la información del botón de programación de funciones en la siguiente línea de texto del archivo de texto anterior:



Para encadenar funciones, como en este ejemplo, sustituir **ENTRADA (CARPETA)/INPUT (FOLDER)** por la salida de la función anterior (C:\EpiMolBio Input/Resultado\);.

Sustituir la salida por otra carpeta. En este caso, generaremos una carpeta llamada “Resultado 2”: **C:\EpiMolBio Input/Resultado2\;**



Se pueden encadenar todas las funciones que se requieran siguiendo este proceso, introduciendo como entrada la salida del proceso anterior.

Una vez finalizada la creación del archivo de texto, guardar el archivo.

En la función de **Herramientas, Programar Funciones**:

En **entrada**, cargar la carpeta donde esté el archivo con las instrucciones (archivo de texto que se ha guardado).

En el campo **Programar Funciones** hacer clic en la casilla para activar la función.

Pulsar **Calcular**.

Con este ejemplo, se genera un fasta sin inserciones (función 1) y traducido (función 2) localizado en la carpeta “Resultado 2”.

No es necesario encadenar siempre las funciones. También se pueden automatizar funciones independientes. Por ejemplo automatizar las salidas de varios archivos html. En este caso la entrada no será la salida de la línea anterior, sino otra entrada independiente.



EpiMolBio

Análisis de la variabilidad genética

Licencia: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (número de registro 2305114294344)

Desarrollador: Roberto Reinosa

Colaboradora: Paloma Troyano-Hernáez

Coordinación y supervisión: África Holguín

Copyright: Roberto Reinosa Fernández y Fundación para la Investigación Biomédica del Hospital Universitario Ramón y Cajal (FIBioHRC)



Fundación para la Investigación Biomédica
del Hospital Universitario Ramón y Cajal



HIV MOLECULAR
EPIDEMIOLOGY
LABORATORY
Instituto Ramón y Cajal de
Investigación Sanitaria (IRYCIS)
Hospital Ramón y Cajal

INSTITUTO
RAMÓN Y CAJAL DE
INVESTIGACIÓN SANITARIA



Hospital Universitario
Ramón y Cajal