Manual de usuario GCP-Shrimp

Juan Sebastián Ramírez Artiles
juan.ramirez@fpct.es
juanseraar@hotmail.com

4 de marzo de 2023

Capítulo 1

GCP-Shrimp: Manual de usuario

En este manual se describen los comandos y funcionalidades implementadas en la aplicación GCP-Shrimp en su versión 5.1. Esta aplicación está desarrollada a medida para cubrir las necesidades operacionales del proyecto ECUANARIA, más concretamente, fue diseñada para facilitar la gestión de datos y la automatización de procesos en las evaluaciones genéticas del camarón blanco ecuatoriano.

1.1. Comandos implementados

- **load-data**: Comando para la carga y actualización de datos.
- **load-families**: Permite la carga de información de apareamientos.
- **check-pedigree**: Comando para chequear el pedigrí antes o después de la carga.
- **eval**: Comando para la evaluación de los individuos.
- **remove-eval**: Comando para la eliminación de evaluaciones almacenadas.
- **list-individuals**: Comando para listar la información relativa a los individuos.
- **list-evals**: Comando para listar los metadatos de las evaluaciones almacenadas.
- **list-ebvs**: Comando para listar la información de EBVs calculada en las evaluaciones.
- **list-eval-nvce**: Lista las asignaciones de nVCEs y sample_ids utilizadas en una determinada evaluación.
- **show-mates**: Lista la información de apareamientos.
- **describe**: Comando para listar las descripciones de las distintas variables.
- **about**: Muestra información sobre la aplicación.

1.2. Opciones de cada comando

- **load-data**:
  - **DATA_FILE**: Nombre del fichero de datos. Es un parámetro obligatorio. El fichero debe estar situado en la carpeta definida en la variable de configuración ***input_data_files_dir***.
  - –**update**: Si está presente, activa la opción de carga con actualización. Los individuos nuevos se cargarán del modo habitual y los duplicados se actualizarán.
  - –**deviation** *FLOAT*: Opción para definir el grado de desviación de los datos considerados fuera de rango. El rango de valores aceptados va de 1 hasta 2. Si no se especifica esta opción se toma automáticamente el valor 1.5.
  - –**check-pedigree**: Activa el chequeo de la genealogía en el proceso de carga. Durante este proceso se validarán y cargará la información de parentesco y los factores fijos que concuerden con la información almacenada de apareamientos. En este proceso se leerá una hoja de Excel con la información de genealogía y se contrastará con la información de los datos de entrada y de los apareamientos. Los apareamientos de la generación anterior a los individuos a cargar debe de haberse cargado previamente en la base de datos. El orden correcto de carga es: primero cargar la generación F0, luego los apareamientos de la F0, y en la carga de los individuos de la F1 ya se podrá realizar la carga con chequeo de pedigrí. Esta opción se puede omitir y la carga se realizará únicamente usando los datos de individuos.
  - –**in-depth**: Parámetro opcional. Si se activa esta opción los individuos a cargar deberán coincidir con los almacenados en la información de apareamientos. En caso de que no se active esta opción la información de nuevos individuos podrá contener más ejemplares que la de genealogía y en el chequeo se usará únicamente la información de los individuos que se encuentre en la genealogía. Los individuos que no estén presentes en la información de apareamientos, se cargarán en la base de datos, pero su información de parentesco y los factores fijos relativos a los apareamientos se asignarán como nulos.
  - –**comment** *TEXT*: Permite introducir información relevante relativa a la carga que se vaya a realizar. El comentario debe de introducirse entre comillas dobles.
  - –**help**: Muestra la ayuda de este comando.
- **load-families**:
  - –**update**: Activa la actualización de la información de apareamientos.
  - –**comment** *TEXT*: Permite introducir información relevante relativa a la carga que se vaya a realizar. El comentario debe de introducirse entre comillas dobles.
  - –**help**: Muestra la ayuda de este comando.

- **check-pedigree**:
  - –**on-database**: Parámetro opcional. Si se activa se realizará el chequeo de la genealogía sobre la información de individuos almacenada en la base de datos. Si el chequeo finaliza correctamente se actualizarán los registros de individuos con la nueva información de parentesco y de factores fijos. Si el chequeo finaliza con incidencias se generará un registro de seguimiento con las incidencias. Independientemente de si se superó o no el chequeo, se le preguntará al usuario si desea hacer los cambios en la base de datos o dejarla sin modificar.
  - –**help**: Muestra la ayuda de este comando.
- **eval**:
  - –**factors** *TEXT*: Permite establecer los factores a evaluar en la propia orden. Los factores deberán pasarse como una cadena de texto separada por comas y sin espacios. En las evaluaciones normales se deberá indicar como mínimo dos factores aleatorios y un factor fijo, mientras que en las evaluaciones con interacción genotipo ambiente se deberá indicar como mínimo un factor aleatorio y un factor fijo. Esta opción prevalece sobre la opción –**select-factors**.
  - –**select-factors**: Si esta opción está presente se mostrará un menú con los factores aleatorios y los factores fijos disponibles en la base de datos. El usuario podrá seleccionar los factores que desee usar en la evaluación. La cantidad de factores mínima será la misma que en el caso de la opción –**factors**.
  - –**generations** *TEXT*: Esta opción permite seleccionar la o las generaciones de individuos que se deseen evaluar. Las generaciones seleccionadas deberán pasarse como una cadena de texto separada por comas y sin espacios. Las generaciones no tienen por que ser consecutivas. Si esta opción no está presente se tomará por defecto la última generación almacenada.
  - –**nulls-rf-allowed** *INT*: Esta opción permite indicar la cantidad máxima de factores aleatorios nulos por registro en el conjunto de datos a filtrar. Si no se incluye esta opción se tomará el valor que esté establecido en la variable de configuración ***default_rf_nulls***.
  - –**nulls-ff-allowed** *INT*: Esta opción permite indicar la cantidad máxima de factores fijos nulos por registro en el conjunto de datos a filtrar. Si no se incluye esta opción se tomará el valor que esté establecido en la variable de configuración ***default_ff_nulls***.
  - –**min-status-one** *INT*: Indica la cantidad mínima de estatus uno que se deberá alcanzar para que la evaluación finalice con éxito. Si esta opción no está presente se tomará el valor que se indique en la variable de configuración ***min_status_one***.
  - –**min-parents** *INT*: Indica la cantidad mínima de padres conocidos por cada individuo. La variable de configuración ***default_min_parents*** establecerá el valor por defecto.

- –**max-combinations** *INT*: Indica el número máximo de combinaciones de variables que se podrán probar en las evaluaciones. Si no se incluye esta opción se tomará el valor por defecto establecido en la variable de configuración ***max_combinations***.
- –**ancestors**: Si esta opción está presente, se añadirá al conjunto de datos la información de los ancestros de los individuos a evaluar. Si no se incluye esta opción, no se añadirán.
- –**interaction**: Si esta opción está presente la evaluación será de interacción genotipo ambiente.
- –**sample-types-excluded** *TEXT*: Permite establecer un filtrado excluyendo a los tipos de muestreo que se pasen en la opción. Opcionalmente permite un listado separado por comas de los distintos tipos de muestreo.
- –**stations** *TEXT*: Permite seleccionar únicamente las estaciones que se pasen como parámetro. Acepta una lista de estaciones separadas por comas. Si esta opción no se incluye en la orden, se añadirán todas las estaciones.
- –**max-depth** *INT*: Establece el nivel de profundidad del pedigrí: 1 para padres, 2 para abuelos, etc. Por defecto toma el valor de la configuración.
- –**num-cores** *INT*: Establece el número de núcleos máximo a usar en la evaluación. Por defecto tomará el valor de la configuración.
- –**help**: Muestra la ayuda de este comando.

▪ **remove-eval**:

- **EVAL_CODE**: Código único de la evaluación a eliminar. Este parámetro es obligatorio. Al eliminar la evaluación se eliminará la meta-información de la evaluación y todos los EBVs calculados en la evaluación.
- –**help**: Muestra la ayuda de este comando.

▪ **list-individuals**:

- –**generations** *TEXT*: Esta opción permite seleccionar la o las generaciones de individuos que se desee listar. Las generaciones seleccionadas deberán pasarse como una cadena de texto separada por comas y sin espacios. Si esta opción no está presente se listarán los individuos de todas las generaciones.
- –**fields** *TEXT*: Permite seleccionar los campos a listar de información relativa a los individuos. La lista de campos deberá pasarse como una cadena de texto separada por comas y sin espacios. Existen tres valores especiales de campos: *meta*, *random*, y *fixed*, que permiten añadir todos los campos de información relativos a la meta-información, a los factores aleatorios, o a los factores fijos. Si esta opción no está presente, se listan todos los campos.
- –**ord-fields** *TEXT*: Acepta una cadena de texto con los campos separados por comas y sin espacios que se quieran utilizar como criterio de ordenación. Si no se incluye esta opción, el orden del listado que se mostrará será el de inserción en la base de datos.

- –**asc**: Permite ordenar el listado de modo ascendente según los criterios de ordenación definidos en la opción –**ord-fields**. Si no se incluye, el orden será descendente.
- –**csv**: Permite listar y exportar a fichero CSV. Si no se incluye la opción –**file-name**, el nombre por defecto será *individuals.csv*.
- –**file-name** *TEXT*: Permite establecer un nombre de fichero para la exportación a CSV. El fichero se generará en la carpeta definida en la variable ***output_files*** del fichero de configuración.
- –**help**: Muestra la ayuda de este comando.

▪ **list-evals**

- **EVAL_CODES**: Parámetro opcional que permite seleccionar un conjunto de evaluaciones a listar. El comando acepta una cadena de texto separada por comas con los códigos únicos de las evaluaciones que se deseen listar. Si esta opción no está presente se listarán todas las evaluaciones.
- –**help**: Muestra la ayuda de este comando.

▪ **list-ebvs**:

- **EVAL_CODE**: Parámetro obligatorio. Código único de la evaluación de la cual se desee listar sus EBVs.
- –**factors** *TEXT*: Opción para seleccionar los EBVs de factores aleatorios de los empleados en la evaluación. Adicionalmente se podrán seleccionar las varianzas de estos factores en columnas aparte.
- –**varians**: Opción que activa la visualización conjunta de los EBVs de cada factor con la variaza calculada en el proceso de evaluación. Si esta opción no está presente no se mostrarán las varianzas junto con los factores.
- –**ord-factors** *TEXT*: Acepta una cadena de texto con los factores separados por comas y sin espacios que se quieran utilizar como criterio de ordenación. Adicionalmente, se podrá ordenar por valores de varianza. Si no se incluye esta opción, el orden del listado que se mostrará será el de inserción en la base de datos.
- –**asc**: Permite ordenar el listado de modo ascendente según los criterios de ordenación definidos en la opción –**ord-factors**. Si no se incluye, el orden será descendente.
- –**file-name** *TEXT*: Permite establecer un nombre de fichero para la exportación a CSV. El fichero se generará en la carpeta definida en la variable ***output_files*** del fichero de configuración.
- –**csv**: Permite listar y exportar a fichero CSV. Si no se incluye la opción –**file-name**, el nombre por defecto será *ebvs.csv*.
- –**environ** *TEXT*: Si los EBVs a listar se obtuvieron en una evaluación con interacción, esta opción permitirá seleccionar el entorno a listar. Si no se indica, se listarán los EBVs de todos los entornos.
- –**help**: Muestra la ayuda de este comando.

- **list-eval-nvce** *INT*:
  - –**file-name** *TEXT*: Permite establecer un nombre de fichero para la exportación a CSV. El fichero se generará en la carpeta definida en la variable ***output_files*** del fichero de configuración.
  - –**csv**: Permite listar y exportar a fichero CSV. Si no se incluye la opción –**file-name**, el nombre por defecto será *nVCEs.csv*.
  - –**help**: Muestra la ayuda de este comando.
- **show-mates**:
  - –**generations** *TEXT*: Admite una lista separada por comas de las generaciones que se desee listar. Si no se incluye esta opción se listarán todas las generaciones.
  - –**fields** *TEXT*: Parámetro opcional. Admite una lista separada por comas para seleccionar los campos que se desee filtrar. Si no se incluye esta opción se listarán todos los campos.
  - –**ord-fields** *TEXT*: Permite activar la ordenación y seleccionar los campos por los los que ordenar.
  - –**file-name** *TEXT*: Permite establecer un nombre de fichero para la exportación a CSV. El fichero se generará en la carpeta definida en la variable ***output_files*** del fichero de configuración.
  - –**csv**: Permite listar y exportar a fichero CSV. Si no se incluye la opción –**file-name**, el nombre por defecto será *mates.csv*.
  - –**help**: Muestra la ayuda de este comando.
- **describe**:
  - –**fields** *TEXT*: parámetro optativo para listar únicamente las variables pasadas en la orden. El formato es un listado de nombres de variables separadas comas. En caso de no incluir esta opción, se listarán todas las variables de la base de datos.
  - –**help**: Muestra la ayuda de este comando.

1.3. Estructura de directorios

En esta sección se listarán los directorios de la aplicación y la funcionalidad de cada uno de ellos.

- **conf**: Almacena el fichero de configuración ***config.yaml***.
- **evaluation_input_files**: Contiene los ficheros de entrada generados para cada evaluación. Estos ficheros son ***ddbb.txt***, ***pedigree.prn***, ***variables.txt***, y ***fvariables.txt***.
- **input_data_files**: En este directorio se deberán situar los ficheros datos Excel que se vayan a utilizar en el proceso de carga de nuevos datos a la aplicación. En la versión GUI, el fichero Excel que se vaya a cargar se copiará desde el directorio de origen a este directorio antes de proceder a la carga.
- **logs**: Este directorio almacenará los ficheros de seguimiento de la aplicación. Los ficheros son tres: "error.log", "activity.log", y "outliers.log". Se generarán automáticamente en caso de que no existan y las nuevas incidencias se irán añadiendo al final de cada fichero.
- **output_files**: En este directorio se situarán los ficheros de datos que se generen al volcar los listados de información de individuos y de EBVs.
- **results**: Este directorio se usará para guardar los ficheros generados como resultado de las evaluaciones. Cada subdirectorio contendrá la información de una evaluación y tendrá como nombre la marca de tiempo de su finalización. La marca de tiempo tendrá el formato: YYYYMMDDhhmmss. En estos subdirectorios se guardarán los directorios vce_analysis con los informes de las evaluaciones finalizadas con status uno.
- **temp_evaluation_process**: Este directorio se generará en cada evaluación y será el directorio desde el que se ejecutarán las evaluaciones. A este fichero se copiarán los ficheros de entrada y el ejecutable ***vce5.exe***. Además, en este directorio se generarán los subdirectorios "finalised_races", "races_in_progress", y "reports". Este directorio se generará al inicio de cada evaluación eliminando la información generada en la anterior evaluación.
- **VCE**: En este directorio se almacena una copia permanente de la aplicación ***vce5.exe***.

1.4. Interfaz gráfica

1. Interfaz de los comandos **load-data**, **load-families** y **check-pedigree**.

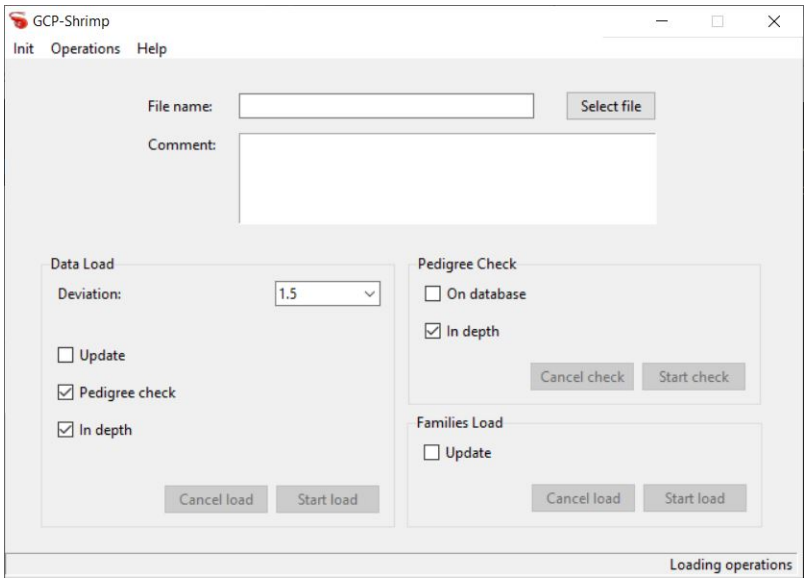

Figura 1.1: Interfaz de carga y chequeo de datos

2. Interfaz del comando **eval**.

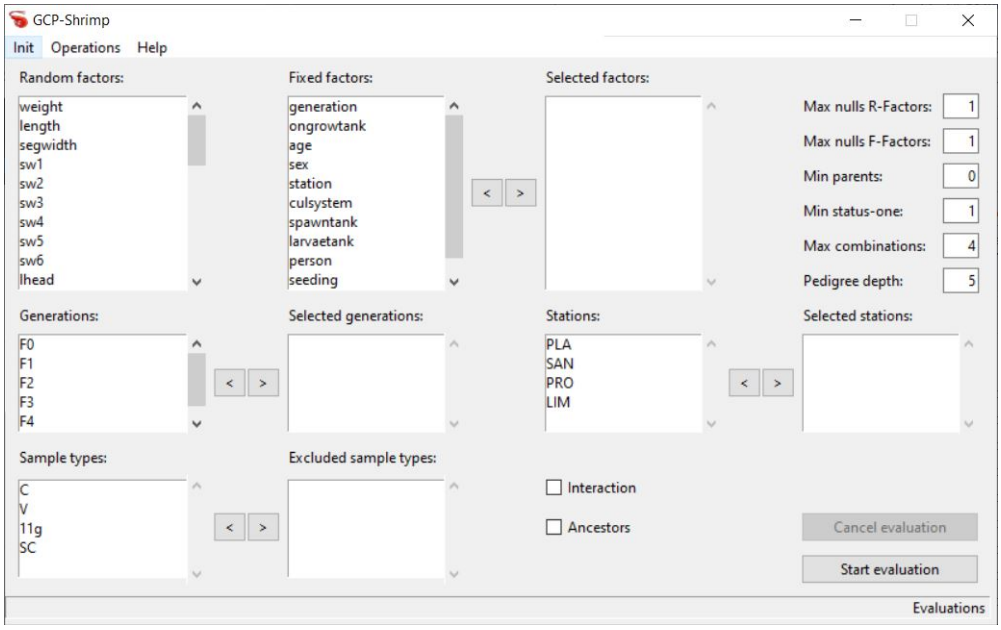

Figura 1.2: Interfaz para la ejecución de evaluaciones

3. Interfaz de los comandos **list-evals** y **remove-eval**.

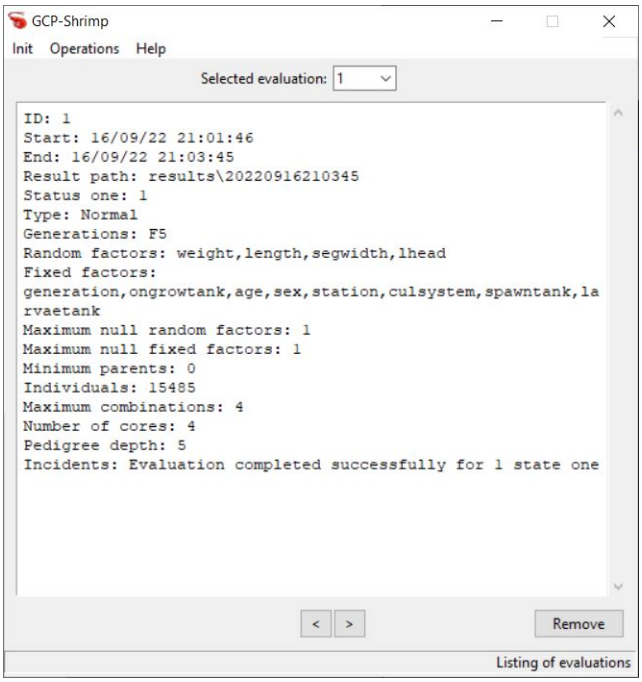

Figura 1.3: Interfaz para listar y eliminar evaluaciones

4. Interfaz para el comando **list-individuals**.

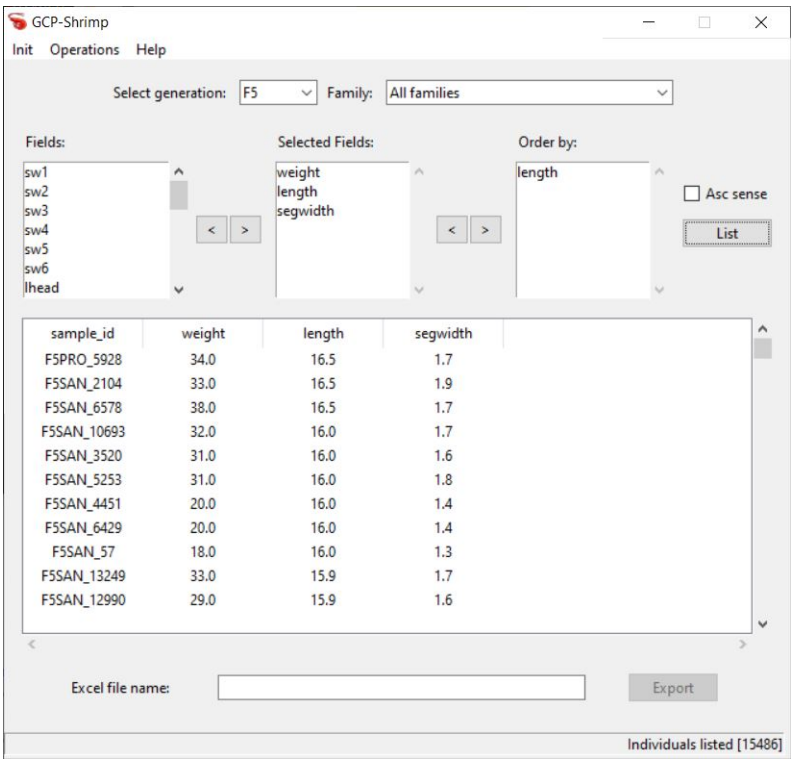

Figura 1.4: Interfaz para listar información de individuos

5. Interfaz para el comando **list-ebvs**.

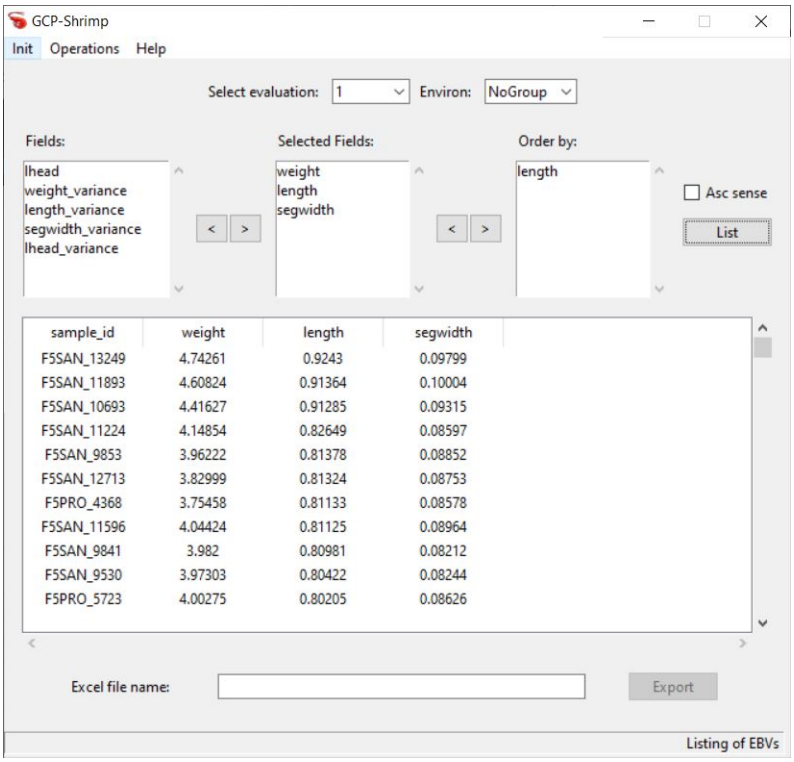

Figura 1.5: Interfaz para listar EBVs

6. Interfaz para el comando **show-mates**.

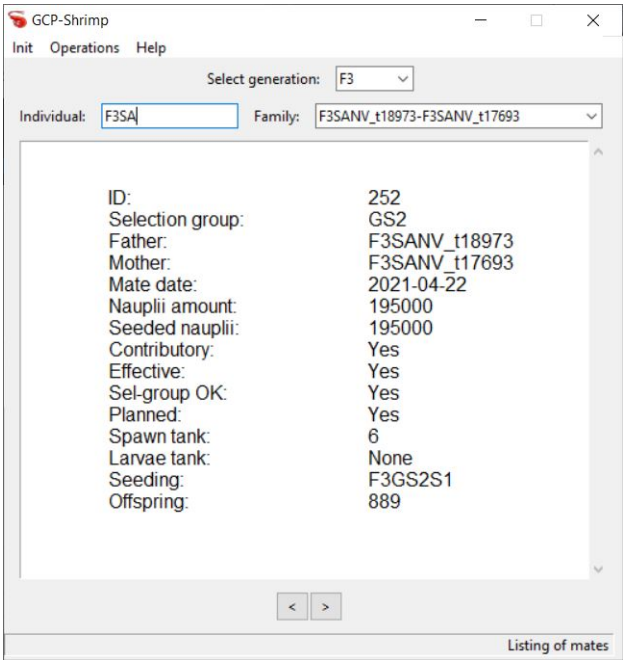

Figura 1.6: Interfaz para listar información de apareamientos

7. Interfaz para el comando **describe**.

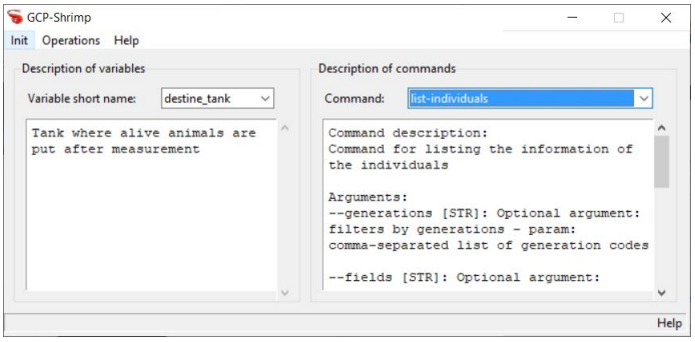

Figura 1.7: Interfaz de ayuda

1.5. Variables y formato de los ficheros de entrada de datos

Como guía de uso de la aplicación vamos a mostrar las variables que usa la aplicación y el formato que deben tener los ficheros de entrada. No todas las variables que trae inicialmente la aplicación son de obligado uso, por este motivo daremos una pequeña descripción de cada variable.

Los datos de entrada deberán pasarse al programa como ficheros Excel. Estos ficheros deberán de contener hojas de datos con nombres predeterminados. El nombre de las distintas hojas se podrá establecer en el fichero de configuración de la aplicación.

Si surgiera la necesidad de crear nuevas variables, simplemente habría que añadir una columna con un nombre nuevo en la hoja de carga de datos. El programa detectará la nueva variable y le preguntará si desea crear una nueva variable, ignorarla o abortar la carga para examinar la corrección de los nombres de los campos por si hubiera un error ortográfico al construir el fichero de entrada.

Las variables de la aplicación se agruparán en tres tipos distintos. Los factores aleatorios, que serán las variables correspondientes a las medidas de morfología; los factores fijos, que serán pasados al modelo BLUP para los ajustes de sesgos en los cálculos; y otras variables, que se calificarán como meta-información.

En el proceso de carga, los datos de código de ejemplar (sample_id), código de estación (station), y generación (generation), son obligatorios. El resto de campos se pueden omitir. El sample_id debe de ser un código único para cada ejemplar dentro de la base de datos.

1.5.1. Factores aleatorias:

- **weight**: Peso medido al tamaño de la cosecha o a la edad específica para el análisis
- **length**: Longitud medido al tamaño de la cosecha o a la edad específica para el análisis
- **segwidth**: Longitud del primer segmento, vivo
- **sw1**: Longitud del primer segmento, descongelado
- **sw2**: Longitud del segundo segmento, descongelado

- **sw3**: Longitud del tercer segmento, descongelado
- **sw4**: Longitud del cuarto segmento, descongelado
- **sw5**: Longitud del quinto segmento, descongelado
- **sw6**: Longitud del sexto segmento, descongelado
- **lhead**: Longitud del cefalotórax desde el rostrum hasta el inicio del primer segmento
- **labdomen**: Longitud del abdomen desde la línea donde comienza el primer segmento hasta la línea donde termina el sexto segmento
- **whead**: Ancho del cefalotórax
- **wabdomen**: Ancho del abdomen
- **hhead**: Alto del cefalotórax
- **habdomen** Alto abdomen
- **cm3**: Volumen del ejemplar
- **cflength**: Factor de condición con el peso y la longitud
- **cfwidth**: Factor de condición con el peso y el ancho del primer segmento
- **deformity**: Presencia o ausencia de deformidad
- **sh1**: Altura primer segmento, descongelado
- **sh2**: Altura segundo segmento, descongelado
- **sh3**: Altura tercer segmento, descongelado
- **sh4**: Altura cuarto segmento, descongelado
- **sh5**: Altura quinto segmento, descongelado
- **sh6**: Altura sexto segmento, descongelado
- **sl1**: Longitud primer segmento, descongelado
- **sl2**: Longitud segundo segmento, descongelado
- **sl3**: Longitud tercer segmento, descongelado
- **sl4**: Longitud cuarto segmento, descongelado
- **sl5**: Longitud quinto segmento, descongelado
- **sl6**: Longitud sexto segmento, descongelado
- **wwmuscle**: Peso húmedo
- **lipids**: Porcentaje de lípidos

- **protein**: Porcentaje de proteínas
- **mineral**: Porcentaje de minerales
- **moisture**: Porcentaje de humedad
- **unfrzleng**: Longitud descongelado
- **sv1**: Volumen del primer segmento
- **sv2**: Volumen del segundo segmento
- **sv3**: Volumen del tercer segmento
- **sv4**: Volumen del cuarto segmento
- **sv5**: Volumen del quinto segmento
- **sv6**: Volumen del sexto segmento

1.5.2. Factores fijos:

- **generation**: Código ordinal de la generación del ejemplar
- **ongrowtank**: Tanque de cría hasta la edad de recolección. Dependiente de la generación.
- **age**: Edad en días del animal
- **sex**: Sexo: masculino (M), femenino (F)
- **station**: Estación de cría hasta la edad de recolección
- **curlsystem**: Sistema de cultivo: Ej.: alta o baja salinidad.
- **spawntank**: Tanque de desove. Dependiente de la generación.
- **larvaetank**: Tanque de larvario. Dependiente de la generación.
- **person**: Nombre de la persona que tomó las medidas
- **seeding**: Número ordinal de siembra. Dependiente de la generación.
- **selgroup**: Grupo de selección de apareamientos. Dependiente de la generación.

1.5.3. Meta información:

- **sample_id**: Código del ejemplar
- **measured_date**: Fecha de toma de medición
- **destine_tank**: Tanque de cría del espécimen
- **ring_code**: Color y número de la etiqueta del espécimen
- **sample_num**: Número de muestra

- **observations**: Cualquier tipo de dato relevante
- **survival**: Vivo (1), muerto (0), sacrificado (null)
- **selected**: Seleccionado para apareamiento (1), no selecionado (0), desconocido (null)
- **sample_type**: Tipo de muestreo: durante sacrificio, en el momento de recolección o en otros momentos
- **father_code**: Código de ejemplar del padre
- **mother_code**: Código de ejemplar de la madre

1.6. Ejemplos de comandos

```
autopro.exe load-data Ejemplo_F2.xlsx --update
```

Figura 1.8: Carga con actualización

```
autopro.exe eval --factors weight,length,age,sex --nulls-allowed 0 --generations F1,F2
```

Figura 1.9: Evaluación con filtrado de factores en la orden

```
autopro.exe eval --select-factors --nulls-allowed 0 --generations F1,F2 --min-status-one 3 --interacion --ancestors
```

Figura 1.10: Evaluación con interacción y ancestros

```
autopro.exe list-ebvs 3 --var --factors cm3,hhead --ord-factors cm3 --asc --csv --file-name cm3_hhead.csv
```

Figura 1.11: Listado de EBVs con ordenación y exportación a CSV

```
autopro.exe list-individuals --fields weight,cm3,meta --csv --generations F1,F2 --file-name weight_cm3_meta.csv
```

Figura 1.12: Listado de individuos con exportación a CSV

```
autopro.exe list-individuals --fields generation,weight,cm3,meta --ord-fields generation,cm3 --asc --generations F1,F2
```

Figura 1.13: Listado de individuos con ordenación