DESARROLLO DE UN SOFTWARE PARA SIMULAR LA VARIACIÓN DE LOS ESTADÍSTICOS F DE WRIGHT EN POBLACIONES CON DIFERENTES TÁCTICAS REPRODUCTIVAS Y TASAS DE DISPERSIÓN DE MACHOS Y HEMBRAS.

OSCAR SUESCUN

TRABAJO DE GRADO
Presentado como requisito parcial
Para optar al título de

Biólogo

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE CIENCIAS
CARRERA DE BIOLOGÍA
Bogotá, D. C.
Fecha: Junio 21

DESARROLLO DE UN SOFTWARE PARA SIMULAR LA VARIACIÓN DE LOS ESTADÍSTICOS F DE WRIGHT EN POBLACIONES CON DIFERENTES TÁCTICAS REPRODUCTIVAS Y TASAS DE DISPERSIÓN DE MACHOS Y HEMBRAS.

OSCAR SUESCUN

APROBADO

_____

Manuel Ruiz García

Director

_____ _____

Ricardo Orjuela — Jurado

Diana Alvarez — Jurado

DESARROLLO DE UN SOFTWARE PARA SIMULAR LA VARIACIÓN DE LOS ESTADÍSTICOS F DE WRIGHT EN POBLACIONES CON DIFERENTES TÁCTICAS REPRODUCTIVAS Y TASAS DE DISPERSIÓN DE MACHOS Y HEMBRAS.

OSCAR SUESCUN

APROBADO

_____	_____
Ángela Umaña, M Phil.	Andrea Forero
Decano Académico	Director de Carrera

TABLA DE CONTENIDOS

pág.

TABLA DE FIGURAS

pág.

TABLA DE TABLAS

pág.

TABLA DE ANEXOS

pág.

1. Introducción

Mantener la diversidad genética de las poblaciones biológicas es una meta importante en conservación, y para comprender cómo se puede hacer esto es preciso conocer los patrones de flujo genético en las mismas. Cuando el flujo genético entre las subpoblaciones o subgrupos en los que está dividida una población es alto, ocurre una homogenización de la variación genética. Si el flujo es bajo, la deriva genética, selección y mutación en los grupos separados pueden llevar a la diferenciación genética.

Sin embargo, previos estudios que se han enfocado en la influencia del intercambio desigual de individuos a lo largo de la población han usado componentes de varianza genética que no se aplican directamente en poblaciones donde se da organización social. En la organización social, en mamíferos, por ejemplo, se pueden dar grupos de hembras filopátricas que se reproducen con un mismo macho poligínico que ha llegado por migración de otro grupo. En aves esta condición se invierte, es decir, las hembras constituyen el género migratorio. Estas complicaciones implican una dificultad para evaluar la partición de la diversidad genética a diferentes niveles de una población con estructura social, lo cuál interfiere con los esfuerzos de conservación.

Se propone el desarrollo de una herramienta informática que evalúe los índices de fijación, o estadísticos F, en poblaciones con estructura social, y que permita tomar decisiones desde el punto de vista de la conservación.

2. Marco Teórico y Revisión de Literatura

La genética de poblaciones es el estudio de la distribución de frecuencias alélicas y su cambio bajo el efecto de varios factores evolutivos genéticos en una o varias poblaciones. Estos factores pueden ser la selección natural, la deriva genética, la mutación o el flujo genético. Por su lado, las poblaciones están a menudo subdivididas en unidades más pequeñas debido a factores geográficos, ecológicos o etológicos. En este trabajo se hace énfasis en distintos modelos de subdivisión por motivos etológicos, que son los más observados en mamíferos y aves (Hedrick 2005).

Para distinguirlo de la migración y de la dispersión, el **flujo genético** se entiende como movimiento de individuos que representan un aporte reproductivo, entre grupos, resultando en intercambio genético. El flujo genético es crucial para entender mecanismos y potencialidades evolutivas en varias áreas de la genética de poblaciones, por ejemplo el potencial de movimiento de transgenes, o de invasividad de especies no-nativas, o de rescate genético, también conocido como restauración genética (Howard *et al.,*2004; Vilá *et al.,*2003; Gaskin and Schaal, 2002).

La **estructura poblacional** depende de la presencia de subestructuras, es decir, diferencias en la variación genética entre las partes constituyentes de la población, ya sea por deriva genética en una parte localizada de la población, por intercambio desigual de individuos a lo largo de la población, o por selección con diferentes efectos en diferentes partes de la población. Esta representación jerárquica se usa para describir relaciones generales de poblaciones de un organismo y para documentar el patrón espacial de la variación genética (Hedrick 2005).

Sewall Wright, uno de los fundadores de la genética de poblaciones, al estudiar los aspectos genéticos de la estructura poblacional, estudió los efectos del apareamiento

aleatorio y la endogamia en la variedad genética. Ya que de ninguna población natural se puede decir que sea totalmente panmícitica o que se encuentre en un estado de total fijación, propuso expresiones matemáticas que describen los efectos relativos de dos coeficientes complementarios en las frecuencias genéticas. Los coeficientes eran F, o coeficiente de fijación, y P, o coeficiente de panmixia. Los definió como complementarios, de manera que P = 1 - F. Al tratar de describir la pérdida de heterocigocidad de poblaciones reproductivas, Wright inventó un método llamado de los coeficientes de camino. Bajo este método, propuso la cantidad F como un coeficiente de endogamia que indica el alejamiento de la cantidad de homocigocidad bajo apareamiento aleatorio hacia la homocigocidad completa (Wright 1922a). Este coeficiente se podía usar para describir las propiedades de una población, en relación con las de una cepa de reproducción aleatoria, o con las de una cepa fundadora.

Al aplicar análisis de pedigríes de los coeficientes de F a lo largo de la historia reproductiva de una cepa de vacas con alta endogamia (Wright & McPhee 1925), observó que tres fenómenos diferentes podían afectar el valor del coeficiente de endogamia. Uno era la subdivisión de una población en varias subpoblaciones, cada una de las cuáles es panmíctica por dentro, pero aisladas entre sí. Otro era el apareamiento frecuente entre parientes cercanos sin separación en subpoblaciones. En estos dos casos, diferentes escalas de medición de F darían diferentes resultados. El tercer fenómeno que afectaba, era el tamaño pequeño de la población, sobre todo si esta venía de una línea aún menor en tamaño. Al tratar de distinguir estos fenómenos, usó una concepción jerárquica de F, dividida en partes para describir mejor la estructura de la población, según si se refiere a la población total (T), a las subpoblaciones (S) o a los individuos (I). Así, propuso un índice promedio de fijación de los individuos con respecto a las subpoblaciones de una población (F_{IS}), como la rata de la heterocigosidad observada con respecto a la heterocigocidad promedio de las subpoblaciones, y otro índice para la fijación de los individuos con respecto a la población total (F_{IT}). Posteriormente probó (Wright 1943a) que la correlación entre

gametos aleatorios provenientes de la misma subpoblación, relativa al total, se da por la fórmula:

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}}$$

Esta relación entre los tres estadísticos se puede expresar de la siguiente forma:

$$(1 - F_{IT}) = (1 - F_{ST}) \times (1 - F_{IS}) \qquad (1)$$

F_{ST} es una medida de la diferenciación genética entre subpoblaciones y siempre es positivo. F_{IS} y F_{IT} son medidas de la desviación a partir de las proporciones Hardy-Weinberg dentro de las subpoblaciones y en la población total, respectivamente, donde los valores positivos indican una deficiencia de heterocigotos y los valores negativos indican un exceso de heterocigotos (Hedrick 2005). Si hay subpoblaciones primarias (S1) subdivididas en subpoblaciones secundarias (S2), se puede expresar así:

$$(1 - F_{IT}) = (1 - F_{IS1}) \times (1 - F_{S1S2}) \times (1 - F_{S2T})$$

Y así sucesivamente para cualquier grado de divisiones jerárquicas (Wright 1951).

Cockerham (1969), al analizar la distinción que existe entre el efecto de la endogamia sobre la heterocigocidad y sobre la varianza de las frecuencias genéticas en poblaciones finitas, describió los componentes de varianza entre frecuencias genéticas para genes neutrales de individuos. Estas varianzas son: σ_a, σ_b, y σ_w, donde 'a' se refiere a grupos, 'b' a individuos y 'w' a dentro de individuos, de manera que σ_a es la varianza de los grupos, σ_b la de individuos, y σ_w la que hay dentro de los individuos. Con base en estas, Cockerham definió las correlaciones entre las

frecuencias de genes de diferentes individuos en el mismo grupo ($\bar{\theta}$); entre genes dentro de individuos aleatorios de diferentes grupos (F); y entre genes dentro de individuos dentro de grupos ($f = (F - \bar{\theta})/(1 - \bar{\theta})$). $\bar{\theta}$ es la coancestralidad entre individuos del grupo, y F se usa en el mismo sentido que Wright la usó, es decir, como endogamia. También halló relaciones para F y $\bar{\theta}$ dependiendo del sistema de apareamiento de la población. Cuando existe un sistema que evita el apareamiento de los parientes, $F<\bar{\theta}$; para poblaciones monoecias, $F=\bar{\theta}$; y para sistemas para los cuáles las parejas estan más emparentadas que lo aleatorio, $F>\bar{\theta}$. Así mismo, estos sistemas de apareamiento poblacional afectan a los valores de los componentes de varianza y de las correlaciones intraclase, siendo ambas positivas en el primer caso, iguales a cero en el segundo, y negativas en el tercero. Finalmente, en el mismo artículo, se ilustró la manera de estimar estos parámetros en poblaciones grandes y compuestas de subdivisiones de iguales tamaños, y la manera de probar hipótesis acerca de estos valores en terminología de chi-cuadrado y de prueba-F.

En un artículo posterior (Cockerham 1973), extendió el uso de estas fórmulas para el caso de poblaciones pequeñas y divididas en subpoblaciones desiguales, y esclarece algunas concepciones derivadas de su trabajo anterior. Un ejemplo es la distinción entre la definición de Wright de las correlaciones de genes en términos de las que se dan entre gametos, y la que aplica en el análisis de Malécot (1948) en términos de probabilidades de identidad por descendencia. La distinción es que el último usa los promedios de todas las correlaciones entre gametos. El resultado es el mismo cuando las correlaciones se hacen entre genes dentro de un mismo individuo, pero es distinto cuando se hacen las correlaciones entre individuos de un grupo. El promedio de correlaciones de genes entre individuos distintos a menudo es diferente del promedio de las correlaciones de genes en una muestra de gametos de ellos. Cockerham, utilizando un desarrollo de la equivalencia en aplicación de las covarianzas y de las correlaciones intraclase, re-expresa la correlación entre genes dentro de individuos dentro de subpoblaciones, como f, y aclara que para todos los propósitos, $F=F_{IT}$,

$f=F_{IS}$ y $\overline{\theta}=F_{ST}$; donde f es una función de las otras dos (F y $\overline{\theta}$). Igualmente, re-expresó una covarianza (la de genes en diferentes subpoblaciones), como la correlación $\overline{\theta}_g$, cuya estimación requiere información sobre las divisiones no-relacionadas entre las cuáles los genes no están correlacionados, es decir, sobre las jerarquías más separadas. Ya que el análisis de Wright de estas correlaciones es implícito que las correlaciones individuales o intergrupales que se hacen relativas a la población total consideran a la población fundadora (ancestral y carente de subdivisiones) como la total, y en la mayoría de los casos se carece de la información procedente de esta población, $\overline{\theta}_g$ no es estimable.

De manera que las correlaciones estimables (F, $\overline{\theta}$ y $\overline{\theta}_g$) forman a las correlaciones f, F' y $\overline{\theta}'$, donde la prima indica que estas son las versiones estimables de los parámetros (ya que en el tratamiento previo, no estaban definidas en relación a una población total bien definida), y $f = f'$. En estos términos, los estadísticos-F, aplicados a datos provenientes de la presente generación, sin asumir pedigríes individuales, selección o migración pasados, se pueden expresar de la forma:

$$\overline{\theta'} = \frac{\overline{\theta} - \overline{\theta}_g}{1 - \overline{\theta}_g};\ F' = \frac{F - \overline{\theta}_g}{1 - \overline{\theta}_g};\ f' = \frac{F' - \overline{\theta'}}{1 - \overline{\theta'}} = \frac{F - \overline{\theta}}{1 - \overline{\theta}} = f$$

Cockerham (1973) además extendió el uso de componentes de varianza para incluir aislados dentro de subpoblaciones, y agregados de áreas dentro de las cuales se agrupan las subpoblaciones. Es decir, mostró que la diversidad genética de la población total se puede particionar en sus componentes (esto es, dentro de subpoblaciones y entre subpoblaciones), cuando la diversidad genética se define como la heterocigocidad (frecuencia de heterocigotos) esperada bajo equilibrio Hardy-Weinberg. De esta forma se puede aplicar el análisis a la diversidad genética para números grandes de loci entre un número finito de subpoblaciones. Esto

distingue el tratamiento de Cockerham del de Wright, que asume implícitamente un número infinito de subpoblaciones (Cockerham 1973).

Nei (1977) introdujo el método de los promedios ponderados a la estimación de los estadísticos-F al considerar en mayor detalle la naturaleza de los datos primarios desde los cuáles se han de estimar. A menudo se tienen frecuencias alélicas para los individuos de varias subpoblaciones, en varios loci, para alelos múltiples. Nei observó que los estadísticos-F no necesitan definirse como la correlación de gametos en unión sino que se pueden definir como una función de las heterocigocidades observada y esperada. Mostró cómo las frecuencias genotípicas y alélicas se usaban para estimar los estadísticos-F para el caso dialélico, que fue el que consideraron Wright y Cockerham, y mostró cómo estas funciones se extendían al incluir múltiples alelos, expresables exclusivamente en términos de los promedios ponderados (para todas las subpoblaciones) de las heterocigocidades observada y las heterocigocidades esperadas para las subpoblaciones y para la población total. El factor de ponderamiento para cada subpoblación es la relación entre el tamaño de esa subpoblación y el de la población total. Los estadísticos-F, en este tratamiento, son expresables mediante las siguientes ecuaciones:

$$F_{IS} = \frac{H_S - H_I}{H_S};\ F_{ST} = \frac{H_T - H_S}{H_T};\ F_{IT} = \frac{H_T - H_I}{H_T}$$

Donde $H_I = H_O$, es decir, la heterocigosidad observada de los individuos. Esta definición de los estadísticos-F es compatible con la de Cockerham en que no asume presencia o ausencia de selección, ya que se basa en frecuencias de la generación presente, y con la de Wright en que aunque F_{IS} y F_{IT} (que miden las desviaciones de las frecuencias genotípicas en relación con las proporciones de Hardy-Weinberg en las subpoblaciones y en la población total, respectivamente) pueden ser positivos o negativos, F_{ST} (que mide el grado de diferenciación genética entre subpoblaciones) solo puede ser positivo, ya que $H_T >= H_S$. La diferencia es que la definición de Nei

toma la población actual en la que se muestrean las frecuencias genéticas como la total. Un problema de esta definición es que el error de estimación asociado a la varianza debida a la deriva genética puede fluctuar de generación a generación, especialmente en tamaños poblacionales relativamente pequeños.

En el mismo artículo, Nei discute la relación que existe entre los estadísticos-F y el análisis de diversidad genética. Nei (1977) había definido la diversidad genética para un locus como la heterocigocidad esperada bajo equilibrio Hardy-Weinberg, sin tener en cuenta las frecuencias genotípicas reales en la población. Con esta definición, mostró que $H_T = H_S + D_{ST}$, donde D_{ST} es la diversidad genética interpoblacional. Además llamó G_{ST}, o coeficiente de diferenciación genética, a la relación entre D_{ST} y H_T. Este coeficiente es idéntico a F_{ST} según su definición en Nei (1977). Ya que el análisis de diversidad genética se propuso primariamente para aplicarse al promedio de diversidad genética para grandes números de loci, este aspecto se puede extrapolar a la estimación de los estadísticos-F, obteniendo sus valores promedio para varios loci correspondientes a las diversidades genéticas.

Por otro lado, el análisis de diversidad genética también se puede expandir, de modo similar a como Wright expandió sus estadísticos-F para incluir más jerarquías de subdivisión poblacional, de manera que si se incluyen colonias (C) dentro de las subpoblaciones, la heterocigocidad se expresa así: $H_T = H_C + D_{CS} + D_{ST}$, donde H_C y $D_{CS} \equiv H_S - H_C$ son las diversidades genéticas intracoloniales e intercoloniales, dentro de las subpoblaciones (Nei 1973). Los estadísticos-F de Wright para este caso se pueden escribir como $1- F_{IT} = (1- F_{IC})(1- F_{CS})(1- F_{ST})$, donde F_{IC}, F_{CS} y F_{ST} son $(H_C - H_O)/ H_C$, $(H_S - H_C)/ H_S$ y $(H_T - H_S)/ H_T$, respectivamente. Las diferencias entre ambos análisis están en que las componentes de diversidad genética se expresan como una proporción de la diversidad total, mientras que en los estadísticos-F, las cantidades son relaciones de dos tipos diferentes de diversidad genética. La motivación de ambos análisis también es distinta: la del primero es la estimación de variaciones genéticas inter- e intra-poblacionales con respeto al genoma entero del

organismo en cuestión (por eso aquí es importante usar un número grande de loci); y la del segundo es estimar la relación entre las frecuencias genotípicas en la población total y en las subpoblaciones para un solo locus (Nei 1977).

La terminología alrededor de los estadísticos-F ha sido una fuente de confusión. Nei (1977) comenta que el mismo Wright sugiere que la terminología de F no debería usarse para representar la homocigocidad (la que, en análisis de diversidad genética se expresa como *J*). Esto a pesar de que la homocigocidad esta relacionada con la fijación y con la endogamia. Por otro lado, Cockerham (1984), aunque observa que sus parámetros son equivalentes a los estadísticos-F, mantiene su propia nomenclatura (f, F y θ), pues Nei (1976) afirma que F_{ST} varía con el número de poblaciones observadas, lo cuál le da un carácter más de estadístico que de parámetro. Cockerham asocia su nomenclatura a un tratamiento de los estadísticos en el cuál explícitamente se elimina el efecto de muestreo (número de alelos por locus, número de individuos por población, número de poblaciones).

En el mismo artículo, Cockerham (1984), a través de simulaciones comparando las medias y los errores estándar de varios métodos de calcular θ, concluye que el método de suma ponderada a lo largo de los distintos alelos es el más insesgado. También sugiere que el método de suma ponderada a lo largo de distintos loci es insesgado porque cada loci cuenta como una muestra independiente. También propone fórmulas generalizadas de estimación de los parámetros, que no presuponen nada acerca de los números de las poblaciones, los tamaños muestreales, o las frecuencias heterocigóticas. Sin embargo, la corrección para los tamaños muestreales no afecta cuando las muestreas son exhaustivas (Cockerham 1984).

Nei (1986) recalcó que el método simplificado propuesto por Cockerham (1984) se limitaba al caso de poblaciones de igual tamaño, derivadas todas simultáneamente de una población fundacional a la cuál hacian referencia los estadísticos, y que las subpoblaciones no habían cambiado de tamaño evolutivamente, y que todas estas

asunciones eran muy problemáticas a la hora de considerar poblaciones naturales, además de que tampoco tenían en cuenta la diferencia entre las tasas de dispersión entre diferentes subpoblaciones, cosa que su tratamiento, basado en heterocigocidades, tenía en cuenta, al tratar con las frecuencias genéticas disponibles de las poblaciones presentes. Neel y Ward (1972), en estudios de las distribuciones genéticas de tribus amerindias, habian obtenido resultados que indicaban que se puede esperar un exceso de heterocigosidad en poblaciones que se caracterizan por una dispersión predominante de un sexo. Prout (1981) realizó una demostración formal de este efecto. Sin embargo, los componentes de varianza en los modelos de Prout, así como los usados clásicamente en la computación de los índices de fijación (Rothman, Sing y Templeton 1974; Nei 1977; Wright 1951, 1978; Nei y Chesser 1983) y de parentesco (Malecot 1969; Morton *et al.,* 1971; Lalouel y Morton 1973), pueden no ser directamente aplicables a poblaciones que se caracterizan por tener organización social en unidades sociales o linajes como los que se presentan en mamíferos, donde se pueden dar grupos de hembras filopátricas que se reproducen con un mismo macho poligínico que ha llegado por migración de otro grupo, o en aves, donde estas diferencias entre los sexos se dan inversamente (Chesser 1990).

En vista de esta falta de aplicación al caso de poblaciones con estructura social, es útil la definición de Cockerham (1973) de los índices de fijación como

$$F_{ST} = \frac{\theta_w - \theta_g}{1 - \theta_g} \; ; \; F_{IS} = \frac{F - \theta_w}{1 - \theta_w} \; ; \; F_{IT} = \frac{F - \theta_g}{1 - \theta_g} \qquad (2)$$

Donde θ_w es la correlación de genes dentro de poblaciones, θ_g es la correlación de genes entre grupos y F es el coeficiente de endogamia, o correlación de genes dentro de los individuos. Estos valores pueden extraerse a partir de datos de polimorfismo genético (por SNPs o microsatélites). Además, esta definición tiene implícita una jerarquía de la subdivisión de la población que puede extrapolarse precisamente al caso que se observa en las típicas poblaciones con estructura social, en donde se ven

linajes dentro de la subpoblación, que se toma como la jerarquía mayor, en vez de la población. Para reflejar esta organización, Chesser (1991a, b) usó los índices de fijación F_{LS}, F_{IL} y F_{IS} basados en los definidos por Cockerham, donde F_{LS} es la proporción de la varianza genética encontrada entre linajes dentro de la (sub)población, F_{IL} es la correlación de genes dentro de individuos relativos a aquellos que están dentro del linaje, y F_{IS} es la correlación de genes dentro de individuos relativos a aquellos dentro de la (sub)población (la manera de calcular estos índices se mostrará en la metodología).

Dobson et al. (1998) compararon los resultados del modelo con datos empíricos de alelos de alozimas, y datos de pedigrí y demografía en perritos de las praderas (*Cynomys ludovicianus*), cuya táctica reproductiva se caracteriza por una dispersión masculina virtualmente completa, y obtuvieron una excelente conformidad (Chesser, comunicación personal). Por esta razón resulta útil el modelo, pues los valores de Cockerham pueden monitorearse para ver su dinámica a través del tiempo en poblaciones con organización social, y es precisamente esto lo que se propone modelar la herramienta informática propuesta en este trabajo. Esta es necesaria para modelar la variación transgeneracional de la partición de la variación genética de poblaciones con estructura social para contribuir a una mejor toma de decisiones con miras a la conservación de su diversidad genética.

En la literatura reciente sobre programas computacionales para análisis de datos genéticos en genética de poblaciones solo es posible hallar programas que calculen estadísticos F y parámetros de coancestralidad instantáneos, pero no se encuentran programas que hagan simulaciones de la variación transgeneracional de estos valores. Por ejemplo, el marco de inferencia del paquete multipropósito FSTAT, que computa índices básicos de diversidad genética, riqueza alélica y estadísticos F, entre otros, involucra solamente estimadores momentáneos. Lo mismo ocurre con el paquete multipropósito Genepop. Éste último fue uno de los primeros paquetes integrados de genética de poblaciones, y a pesar de llevar un tiempo sin actualizaciones, y de que

sus funcionalidades estén disponibles en varios otros programas regularmente actualizados, ha adquirido una popularidad suficiente como para que varios otros programas usen el formato de entrada de datos de Genepop (Raymond & Rousset 2005) como el estándar en análisis de genética de poblaciones. Además hay programas (p. ej. Convert, Formatomatic) de conversión que toman el formato de datos de Genepop y lo convierte al formato de otros programas. Esta fluidez de formatos es esencial para que un investigador no se limite a los resultados que obtiene con un programa, sino que pueda efectuar varios análisis de sus datos sin tener que reformatearlos manualmente (Excoffier & Heckel, 2006).

Entre los programas computacionales aptos para el análisis de la subdivisión poblacional, incluyendo estadísticos F, los dos ya mencionados (FSTAT y Genepop) manejan marcadores genéticos multialélicos, es decir, loci para los cuales no se asume un modelo mutacional específico, o para los cuáles la mutación es despreciable. En este caso, las computaciones se basan solamente en frecuencias alélicas. De lo contrario, los diferentes programas asumen diferentes modelos mutacionales. Estos dos programas también son aptos para marcadores genéticos tipo STR (repeticiones cortas en tándem, según sus siglas en inglés). El programa especializado Hickory trabaja con datos de marcadores dominantes tipo AFLP (polimorfismo de longitud amplificada de fragmento, según sus siglas en inglés). Usa estimación bayesiana de los estadísticos F en muestras de una población subdividida, y reporta la distribución posterior de los coeficientes de endogamia y F_{ST}, con la posibilidad de comparar la distribución de diferentes conjuntos de datos (Excoffier & Heckel, 2006).

Cuando las relaciones que componen el modelo de un sistema real son suficientemente simples, se pueden usar soluciones analíticas para obtener información exacta sobre cuestiones de interés del modelo. Sin embargo algunos sistemas son tan complejos que es necesario usar simulaciones que evalúan numéricamente el modelo para obtener estimaciones sobre las características del

modelo matemático. Sin embargo un modelo puede proveer soluciones exactas. Una simulación puede ser estática, cuando representa sistemas en los que el tiempo no juega ningún rol, o dinámica, en el caso contrario. Puede ser determinista, cuando los datos de salida estan determinados por los datos de entrada, o estocásticos, cuando el modelo incluye datos aleatorios. Pueden ser continuos (o análogos), cuando el tiempo varía de manera continua, o pueden ser discretos (o basados en eventos), cuando el tiempo varía de manera discreta (Law & Kelton, 1991)

La interacción humano-computadora es el estudio de la interacción entre usuarios y computadoras. Es un tema interdisciplinario, que relaciona la ciencia computacional con otros campos de estudio e investigación. La interacción ocurre a nivel de la interfaz de usuario, que incluye tanto el software como el hardware. La meta de este campo de estudio es proponer una metodología y proceso para el diseño de interfaces. La metodología de diseño centrado en el usuario se enfoca en las necesidades y limitaciones del usuario, para crear un sistema computacional que se adapte a ellas. La manera de lograr esto es mediante conversaciones entre usuarios, diseñadores e ingenieros, para acoplar los sistemas al tipo de experiencia que el usuario desea, más que a un modelo general de atención, memoria y percepción del usuario (Dix 1991).

El diseño de interfaz de usuario se centra en que la interacción del usuario con un sistema computacional sea intuitiva, más que en la estética del sistema. A nivel de software, la interfaz se conoce como el caparazón. El proceso del diseño gráfico de la interfaz gráfica de usuario establece la apariencia (diseño visual) y sensación (comportamiento de los elementos dinámicos) del caparazón. La interfaz gráfica le permite al usuario interactuar con un computador empleando elementos como punteros, menús, ventanas, íconos, indicadores visuales o controles (elementos gráficos especiales, también llamados *widgets*, que incluyen: botones, deslizadores, pestañas, listas interactivas, barras de herramientas, etc.), junto con etiquetas de texto o navegación de texto para representar la información y acciones disponibles para el usuario. Las acciones a menudo se desempeñan a través de una manipulación directa

de los elementos gráficos. Este tipo de caparazón es más intuitivo que el basado en líneas de comando, el cuál requiere que los comandos sean tecleados, ya que suele tener una curva de aprendizaje bastante inclinada. Es decir que, aunque los comandos pueden ser efectivos para especificar tareas complejas, requieren el uso de cadenas de comando que pueden ser difíciles de aprender (Dix 1991).

3. Formulación del Problema y Justificación

3.1 Formulación del Problema

No existe una herramienta de software que evalúe la influencia de la filopatría-migración de las hembras y los machos, y de la táctica reproductiva, en la partición de la variación genética a través del tiempo, para uso en la toma de decisiones en conservación.

3.2 Preguntas de Investigación

¿Cuál es la influencia de la estrategia reproductiva y las tasas de dispersión de hembras y machos en la partición de la diversidad genética entre las jerarquías de una población estructurada?

3.3 Justificación de la Investigación

Las herramientas de software disponibles sólo calculan estadísticos F momentáneos, pero no se encuentran herramientas que los simulen a través de varias generaciones. Por esto se necesita una herramienta informática que evalúe los índices de fijación, o estadísticos F, a través del tiempo en poblaciones con estructura social, y que permita tomar decisiones desde el punto de vista de la conservación. Ésta permitiría manejar, por ejemplo, tropas de monos aulladores, que se caracterizan por tener estructura social (o cualesquiera otros grupos animales que presenten esta característica), y con base en la información genética obtenida de ellos, asesorar el traslado o retención oportunos del número y género de individuos óptimos en un área determinada para mejorar las probabilidades de su viabilidad genética a largo plazo. La herramienta podría además ser de uso en la conservación de especies amenazadas con importancia económica.

4. Objetivos

4.1 Objetivo general

Desarrollar un software que simule la variación transgeneracional de los parámetros de coancestralidad y de los estadísticos F, para poblaciones (modeladas o muestreadas) con organización social, para diferentes estrategias reproductivas y tasas de dispersión de machos y hembras.

4.2 Objetivos específicos

1. Diseñar, implementar y validar un método de ingreso de datos de poblaciones genéticas.
2. Diseñar, implementar y validar un conjunto de rutinas que calculen los parámetros de coancestralidad a partir de datos biológicos.
3. Diseñar, implementar y validar un conjunto de rutinas que simulen y grafiquen la variación transgeneracional de los estadísticos F a partir de los parámetros de coancestralidad, las tasas de dispersión y el índice de poliginia.

5. Materiales y métodos

5.1 Diseño de la Investigación

Al carecer de recolección de muestras, diseño experimental y tratamiento estadístico, este trabajo de grado carece de diseño de investigación, excepto el diseño de software, cuyas fases se describen en los métodos.

5.1.1 Población de Estudio y Muestra

Este estudio, al ser un desarrollo técnico, carece de población de estudio y muestra. Las muestras que puede tomar la herramienta para su aplicación en un estudio, son las frecuencias genotípicas de cualquier población con estructura social.

5.1.2 Variables del Estudio

Las variables del estudio son la metodología de desarrollo de software y el lenguaje de programación (las frecuencias genotípicas, los índices de variación y diferenciación genética y las constantes poblacionales son las variables de aplicación de la herramienta desarrollada).

5.2 Métodos

Para el desarrollo del software, se usaron la metodología de ciclo de vida y el lenguaje de programación Visual Basic.

El lenguaje de programación determina en buena medida el alcance del software, ya que diferentes lenguajes tienen diferentes fortalezas y debilidades. El lenguaje Visual Basic .NET provee una interfase de usuario intuitiva y fácil de manejar, con un buen manejo de gráficas. Este lenguaje, que es una versión derivada del Visual Basic, es bastante conveniente para el uso del paradigma de programación basado en eventos, en el cuál el flujo del programa esta determinado por acciones del usuario, lo cuál permite un nivel considerable de interactividad. El problema que presenta este lenguaje es su escasa amplitud de plataformas (está restringido a Windows, y no es aplicable a Apple o Linux). Sin embargo, este problema podría ser solucionado en un futuro cercano, con ayuda del proyecto 'Mono', que actualmente esta siendo desarrollado por la comunidad de Internet como un intento de extender el lenguaje de Visual Basic .NET para aplicarlo en Linux. Esto sería ideal ya que Linux esta ganando reputación como la plataforma de código abierto preferida por la comunidad científica internacional para aplicaciones de investigación.

Otra consideración a tomar con Visual Basic .NET es el uso de matrices, ya que el lenguaje no tiene implementadas librerías ni funciones de operaciones para dichos elementos matemáticos. Los estadísticos F requieren operaciones entre vectores y matrices, pero es algo fácilmente manejable aplicando la teoría de Algebra Lineal y deduciendo el resultado como un vector unidimensional, evitando del todo la realización de barridos de posiciones y sumatorias.

Otros lenguajes que se pudieron haber usado, como C++ o Fortran, que tienen un manejo estadístico más robusto y corren a mayor velocidad, fueron descartados por su dificultad de uso y su pobre interfase gráfica. Cabe anotar que una parte esencial del

programa serían las gráficas de los estadísticos F, que resumen una base de datos crudos enorme y difícil de analizar. Otro lenguaje que se pudo haber utilizado fue el Java, que tiene implementada la capacidad de correr en Linux. Sin embargo, el lenguaje de Java se basó en el lenguaje C, lo cuál desde la perspectiva del autor del presente trabajo lo pone en desventaja con el lenguaje Visual Basic, ya que este se basa en el lenguaje Basic, con el cuál se encuentra más familiarizado. Sin embargo, no existe un consenso global acerca de qué lenguaje es mejor, en general, o para propósitos particulares.

La metodología de programación es un conjunto codificado de prácticas que se puede desempeñar repetiblemente para producir software. La metodología escogida para el presente trabajo fue la de ciclo de vida. El ciclo de vida es un proceso de elaboración de sistemas de información y aplicaciones informáticas, cuyo objetivo es definir las actividades a ser ejecutadas en el proyecto, estableciendo un punto de control para tomar la decisión de continuar o no con el proyecto. Involucra varios modelos, entre los cuáles esta el clásico, el de prototipos, y en espiral. El clásico (también llamado en cascada o secuencial) es el modelo en el que se basan los demás, con leves modificaciones, y es el que se usa en el presente trabajo (Contreras, 2006).

Las fases del ciclo de vida clásico son: análisis, diseño, programación, pruebas, documentación, implementación y mantenimiento. Durante la fase de **análisis** se establece el problema que se va a resolver mediante el software y se recopila la información que determina las características y el funcionamiento del sistema. En la fase de **diseño** se produce un modelo que cumpla con los requerimientos entregados en la fase de análisis, y tiene dos enfoques. En el **enfoque operativo, o no-funcional,** se consideran los métodos de ingreso y salida de datos, así como el diseño de pantallas (la forma como el usuario percibe el software desarrollado). El **enfoque computacional, o funcional,** determina los algoritmos que manejan las variables, lo cuál requiere el uso de diagramas de flujo y seudo-código, describiendo el almacenamiento y procesamiento de la información. La fase de **programación**, o de

desarrollo o codificación, es en la que se traduce la información ya procesada a un lenguaje que el ordenador pueda interpretar. En la fase de **pruebas**, o de validación, se busca detectar errores, sean computacionales, de diseño, o de análisis. Puede tratarse de **pruebas funcionales**, en las que se comparan los resultados obtenidos con el software con los resultados esperados y se evalúa el comportamiento del sistema en cuanto a velocidad y rendimiento. Otro tipo de pruebas evalúa la aptitud del software para los usuarios, como por ejemplo las **pruebas de concepto**, en las que los usuarios interactúan con el producto y hacen sus comentarios sobre apariencia, eficiencia, facilidad de uso, funcionalidad, interactividad, claridad, etc. La **documentación** de sistemas de información es el conjunto de información que nos dice qué hacen los sistemas, cómo lo hacen y para quién lo hacen. Se divide en dos partes; una es la **documentación técnica**, que describe cómo fue desarrollado el software, que a su vez se divide en documentación interna (comentarios en el código fuente, que se realizan durante la fase de programación) y el manual técnico. La segunda parte es la **documentación de usuario**, que describe el funcionamiento del software y se dirige al usuario final (esta se realiza después de la programación). La fase de **implementación** es el inicio de la vida útil del sistema desarrollado. Implica estrategias de distribución del software y de entrenamiento de los usuarios. La fase de **mantenimiento**, o soporte, implica un retorno a las fases anteriores para incluir cambios que se hayan hecho evidentes después de la implementación (Contreras, 2006).

Los métodos relativos al **primer objetivo** se describen dentro de la sección de diseño operativo, y los relativos al **segundo objetivo** se describen dentro de la sección de diseño computacional. Los métodos relativos al **tercer objetivo** corresponden a las fases de análisis, diseño y programación del ciclo de vida del desarrollo del software.

5.2.1 Análisis

Esta es la primera fase del ciclo de vida del desarrollo de software. El problema que resuelve el software es el planteado en la sección 3 del presente documento, y la información relevante proviene de Chesser (1991a, b). Se pretende desarrollar un software que desempeñe simulaciones dinámicas, deterministas y discretas de los estadísticos F de poblaciones con estructura social. Las variables de entrada (opcionales y no-opcionales), las de proceso y las de salida se documentan en las tablas 1-6.

Tabla 1 Variables de entrada (condicionales).

Nombre de Variable	Descripción	Tipo de Variable	Rangos permitidos
Formato Genepop	Frecuencias genotípicas para distintos loci de los individuos de las distintas subpoblaciones	Formato Genepop	Descritos en las instrucciones del Formato Genepop (Raymond & Rousset 2005).
m	Número de machos reproductores por linaje. Es un promedio general.	Decimal positivo	*m*>=1
b	El número de hembras en un linaje que se reproducen con cada macho. Es un promedio general.	Decimal positivo	0<*b*<=n

Tabla 2 Variables de entrada (constantes poblacionales)

Nombre de Variable	Descripción	Tipo de Variable	Rangos permitidos
r	Número de alelos en la población (se obtiene a partir del formato Genepop)	Entero positivo	$r>0$
s	Número de linajes en la población (se puede obtener a partir del formato Genepop, o como dato inicial)	Entero Positivo	$s>0$
n	Número de hembras en cada linaje. Es un promedio general.	Decimal positivo	$n>1$
ϕ	Poliginia de la población (no-independencia de las parejas de los machos). Mide la probabilidad que tienen diferentes hembras de aparearse con el mismo macho (puede obtenerse a partir de n, m y b, ó como dato inicial).	Decimal positivo	$0<=\phi<=1$
d_m	Dispersión masculina	Entero	[0,1]
d_f	Dispersión femenina	Entero	[0,1]
g	Número de generaciones a calcular, equivalente a número de tiempos t.	Entero Positivo	$g>0$

Tabla 3 Variables de entrada (parámetros de coancestralidad)

Nombre de Variable	Descripción	Tipo de Variable	Rangos permitidos
α	Coancestralidad o correlación genética de padres de diferentes linajes	Decimal Positivo	0<= α <=1
F	Coancestralidad o correlación genética de genes dentro de individuos al azar	Decimal Positivo	0<= F <=1
θ	Coancestralidad o correlación genética de padres de diferentes linajes	Decimal positivo	0<= θ <=1
θ_{mm}	Coancestralidad de descendientes aleatorios masculinos dentro del mismo linaje	Decimal Positivo	0<= θ_{mm} <=1
θ_{mf}	Coancestralidad de descendientes aleatorios de diferente sexo dentro del mismo linaje	Decimal Positivo	0<= θ_{mf} <=1

Tabla 4 Variables de proceso (Parte 1)

Nombre de variable	Descripción	Tipo de variable	Rangos permitidos
i	Población (o macho, cuando se usa como contador en la sumatoria de la fórmula para calcular ϕ)	Entero	Enteros positivos
k	Alelo	Entero	Enteros positivos
N_T	Número de individuos en la (sub)población total.	Entero	Enteros positivos
N_i	Número de individuos en cada linaje.	Entero	$1< N_i< N_T$
N_{ik}	Número de individuos heterocigóticos para el alelo k en la población i.	Entero	$1< N_{ik}< N_T$; $(N_{ik} + N_{ikk})/ N_i = 1$
N_{ikk}	Número de individuos homocigóticos para el alelo k en la población i.	Entero	$1< N_{ikk}< N_T$; $(N_{ik} + N_{ikk})/ N_i = 1$

Tabla 5 Variables de proceso (Parte 2)

Nombre de variable	Descripción	Tipo de variable	Rangos permitidos
P_{ikk}	Probabilidad de hallar el alelo k en forma homocigótica en la subpoblación *i*	Decimal positivo	$0 <= P_{ikk} <= 1$
p_{ik}^2	Probabilidad de hallar el alelo k en la subpoblación *i*, al cuadrado	Decimal positivo	$0 <= p_{ik}^2 <= 1$
p_{ik}	Probabilidad de hallar el alelo k en la subpoblación *i*	Decimal positivo	$0 <= p_{ik} <= 1$
w_i	Factor de ponderamiento de la subpoblación *i*	Decimal positivo	$0 <= w_i <= 1$
P_{kk}	Probabilidad ponderada de hallar el alelo k en forma homocigótica en la población total	Decimal positivo	$0 <= P_{kk} <= 1$
$\overline{p_k^2}$	Promedio ponderado del cuadrado de las probabilidades de hallar el alelo k en la población total.	Decimal positivo	$0 <= \overline{p_k^2} <= 1$
$\overline{p_k}$	Promedio ponderado de probabilidades de hallar el alelo k en la población total.	Decimal positivo	$0 <= \overline{p_k} <= 1$

Tabla 6 Variables de salida

Nombre de Variable	Descripción	Tipo de Variable	Rangos permitidos
F_{LS}	Proporción de la varianza genética encontrada entre linajes dentro de la subpoblación. Este estadístico compara la variabilidad genética dentro de linajes y entre linajes de la subpoblación, con base en datos de polimorfismo genético.	Lista de Decimales	$0 <= F_{LS} <= 1$
F_{IL}	Correlación de genes dentro de individuos relativos a los del linaje. Es una medida de la desviación de las proporciones de Hardy-Weinberg por endogamia dentro de los linajes.	Lista de Decimales	$-1 <= F_{IL} <= 1$
F_{IS}	Correlación de genes dentro de individuos relativos a los de la subpoblación. Es una medida de la desviación de las proporciones de Hardy-Weinberg por endogamia entre los linajes .	Lista de Decimales	$-1 <= F_{IS} <= 1$

5.2.2 Diseño

Diseño operativo

Para el método de ingreso de datos, el programa valida los datos alertando cuando los datos no están dentro de los rangos apropiados o de la forma decimal apropiada, y pidiendo reemplazar los datos. El ingreso de n, s, m, b y g es por medio de cajas de texto, en donde el usuario teclea los valores deseados. El ingreso de las tasas de dispersión (d_m y d_f) se divide en tres casos, uno para cada matriz de transición usada. Estos casos se seleccionan por medio de botones radiales. Los primeros dos casos asumen valores fijos de d_m y d_f, y el tercer caso requiere el ingreso de valores de d_m y d_f mediante cajas de texto. El ingreso de ϕ se efectúa por selección, mediante botones radiales, de dos casos, uno en el cuál ϕ se calcula a partir de los valores ingresados de n, m y b, y otro caso en el cuál ϕ se ingresa directamente en una caja de texto.

El ingreso de las correlaciones genéticas, o parámetros de coancestralidad (α, F, θ_{mm} y θ_{mf}) es opcional. Puede hacerse introduciendo los valores en cajas de texto, o usando las opciones de Genepop. Las opciones de Genepop permiten, mediante un botón normal, cargar los datos de un archivo de Genepop previamente creado. En el momento de cargar el archivo de Genepop, existe la opción, mediante botones radiales, de calcular w_i mediante dos métodos: uno asumiendo una ponderación equitativa entre linajes, y otro que atribuye ponderaciones distintas. El ingreso de opciones de visualización de la gráfica usa botones de chequeo para seleccionar cualquier combinación de cuatro opciones: usar símbolos en las curvas, usar fondo en la gráfica, mostrar la grilla del eje de las X, y mostrar la grilla del eje de las Y. Cuando todos los anteriores datos han sido seleccionados, se ingresan conjuntamente mediante un botón normal destinado a calcular los estadísticos F.

Para los métodos de salida de datos, el programa posee una modalidad por defecto, que es la generación de la gráfica correspondiente a las variables de entrada, junto con etiquetas para las diferentes curvas y una etiqueta mostrando los valores de las variables de entrada, junto a las curvas. En la segunda pestaña del programa muestra una tabla (que se puede exportar a Excel mediante un botón) con los valores para cada generación (desde la cero hasta la deseada) de las cinco curvas graficadas ($1 - \alpha$, $1 - \theta$, F_{LS}, F_{IL} y F_{IS}). El programa también posee una modalidad opcional, que es la generación de un archivo gráfico independiente, mostrando la gráfica producida, para su uso en presentaciones o simplemente para tener un registro. El archivo gráfico puede tener extensiones (*.emf), (*.png), (*.gif), (*.jpg), (*.tif), o (*.bmp).

Cuando se ha cargado un archivo de Genepop, además de la gráfica, el programa genera tablas de los valores intermedios calculados en una pestaña auxiliar, mostrando las frecuencias genotípicas, las frecuencias alélicas, homocigóticas y heterocigóticas por población, las frecuencias homocigóticas y heterocigóticas por alelo, los loci etiquetados, el número de individuos y valores de w_i por población, los parámetros de coancestralidad momentáneos, y los estadísticos F momentáneos, junto con una comprobación de la regla de Wright a partir de ellos. Las tablas permiten ser copiadas y pegadas como texto normal en tablas de Excel.

Una vez seleccionados unos datos de entrada para el programa, estos se pueden guardar oprimiendo el botón que representa una unidad de memoria (Save). El programa guarda automáticamente los datos de entrada remanentes en el programa antes de haberlo cerrado por última vez, en un archivo localizado en el mismo directorio donde se encuentra el programa, y llamado ‘lastsession.fsm’. Estos datos serán guardados en la carpeta y con el nombre deseados, en formato (*.fsm). Al oprimir el botón que representa una carpeta abriéndose (Open), se puede buscar un documento guardado con el programa, para abrirlo, generándose automáticamente la simulación correspondiente a los datos guardados.

El diseño gráfico mantuvo un enfoque minimalista, usando colores claros para la presentación de las tablas de datos, resaltando de rojo los valores que el usuario debe introducir, y con versatilidad de colores en la presentación de las gráficas para distinguir las diferentes curvas, además del uso opcional de símbolos para distinguirlos incluso en el caso de que no se disponga de colores al mostrar la gráfica salvada, o de un fondo con gradiente gris que resalta estéticamente la gráfica para el caso en el que se dispone de un uso completo de colores para su presentación. Las grillas de las X y Y permiten una evaluación visual más precisa de los diferentes valores en diferentes generaciones. Al momento de ser generada, la gráfica se puede deslizar y redimensionar, además de que los datos precisos de cualquier punto de cualquier curva se pueden consultar llevando el puntero sobre ese punto. El botón que efectúa la simulación se hizo especialmente grande y de un color suave pero distinto y visible, pues es el que desencadena el grueso de las actividades del programa.

El programa corre a gran velocidad y sus ventanas se dejan manipular fácilmente, para expandirlas/contraerlas o arrastrarlas por la pantalla. En computadores de pantalla pequeña es posible que para ver toda la información se requiera el uso de la barra deslizante, pero esto es preferible a tener que usar más pestañas, pues es más intuitivo. Ya que existe otro programa que calcula estadísticos momentáneos F, llamado FSTAT, mientras que éste los calcula para varias generaciones, a modo de simulador, se ha optado por llamar al programa "FStatSim".

Diseño computacional

Los métodos matemáticos descritos a continuación se basan en Hedrick (2005), Nei (1977), Cockerham (1973) y Chesser (1991a, b), donde se definieron y/o describieron las variables, fórmulas y operaciones usadas en este trabajo. Estos procedimientos involucran los datos especificados en las tablas de datos de entrada, de proceso y de salida, mostradas más arriba. A partir de los datos de frecuencias genotípicas,

ingresados opcionalmente en el Formato Genepop, se estiman las frecuencias alélicas y homocigóticas para cada linaje *i*, por la técnica de conteo genético:

$$p_{ik} = \frac{(2 \times N_{ikk}) + N_{ik}}{2 \times N_i}$$
$$p_{ikk} = \frac{N_{ikk}}{N_i} \quad (3)$$

Donde p_{ik} es la frecuencia del alelo *k*, p_{ikk} es la frecuencia homocigótica del alelo *k*, N_{ikk} es el número de individuos homocigotos para el alelo *k*, N_{ik} es el número de individuos heterocigotos para el alelo *k*, y N_i es el número total de individuos del linaje *i*.

A continuación, se calculan los parámetros de coancestralidad (*F*, θ y α) mediante las ecuaciones de Nei:

$$P_{kk} = \sum_{i}^{s} w_i \times P_{ikk}$$
$$\overline{p_k^2} = \sum_{i}^{s} w_i \times p_{ik}^2$$
$$\overline{p_k} = \sum_{i}^{s} w_i \times p_{ik}$$
$$F = (1 - H_0) = J_0 = \sum_{k=1}^{r} P_{kk} \quad (4)$$
$$\theta = (1 - H_S) = J_S = \sum_{k=1}^{r} \overline{p_k^2}$$
$$\alpha = (1 - H_T) = J_T = \sum_{k=1}^{r} \overline{p_k}^2$$

Donde, opcionalmente, w_i se calcula como $w_i = \frac{1}{s}$ cuando se consideran a todas las *s* subpoblaciones como equivalentes en ponderación, o como $w_i = \frac{N_i}{N_T}$ cuando se consideran los linajes como de distinto tamaño relativo, donde N_i es el número de individuos en el *i*-ésimo linaje, y N_T es el número en la (sub)población total. Las *H*

(heterocigosidades)y las *J* (homocigosidades) son las expresiones en la nomenclatura de Nei (en los subíndices, la *O* significa 'observada dentro de un linaje para todos los loci', la *S* significa 'esperada dentro de linajes para todos los loci' y la *T* significa 'esperada en la población total'). A continuación se calculan los siguientes valores usando las variables de entrada:

Poliginia (ϕ)

$$\phi = \frac{\sum_{i=1}^{m}[b_i^2 - b_i]}{n^2 - n} \tag{5}$$

En este caso se toma b_i como equivalente para cada linaje, de manera que se reemplaza por b, que es un promedio general del número de hembras que se aparean con los machos reproductores de cada linaje.

Vector S_t (tiempo inicial es t = 0)

$$S_t = \begin{pmatrix} \alpha_t \\ F_t \\ \theta_{mm_t} \\ \theta_{mf_t} \end{pmatrix} \tag{6}$$

Para el caso por defecto (sin información del formato de Genepop), los valores iniciales de las correlaciones genéticas (S_0) son todos iguales a cero.

Para el caso de la inclusión de información del formato de Genepop, a partir de las frecuencias genotípicas se calculan las correlaciones genéticas α, F y θ, y para obtener θ_{mm} y θ_{mf} se usa la fórmula:

$$\theta = (\theta_{mm} + \theta_{mf}) / 2 \tag{7}$$

De manera que $\theta_{mm} = \theta_{mf} = \theta$.

Matrices de transición:

T_1: cambios probabilísticos de la columna vector de parámetros de coancestralidad para el caso de machos migrantes con hembras filopátricas ($d_m = 1$ y $d_f = 0$).

T_2: cambios probabilísticos de la columna vector de parámetros de coancestralidad para el caso de machos y hembras migrantes ($d_m = d_f = 1$).

T_3: cambios probabilísticos de la columna vector de parámetros de coancestralidad para el caso de machos y hembras con combinaciones de valores de dispersión diferentes de ($d_m = 1$, $d_f = 0$) y ($d_m = d_f = 1$).

Para simplificar la presentación de los términos se definen:

$$A = (d_m + d_f - d_m d_f); \tag{8}$$

$$B = (1 - x)(d_m(1 - \phi) + d_f); \tag{9}$$

$$x = \frac{n-1}{ns-1}; \tag{10}$$

$$y = \frac{1}{s} \tag{11}$$

Y se incluyen en la definición de las matrices:

$$T_1 = \begin{pmatrix} \frac{4-x-2y}{4} & 0 & \frac{x}{4} & \frac{y}{2} \\ 1-y & 0 & 0 & y \\ \frac{2(1-y)+(1-x)(1-\phi)}{4} & \frac{\phi}{8} & \frac{1+x(1-\phi)}{4} & \frac{y}{2} \\ \frac{2n(1-y)+(n-1)(1-x)(1-\phi)}{4n} & \frac{\phi(n-1)+2}{8n} & \frac{(n-1)(1+x(1-\phi))}{4n} & \frac{y}{2} \end{pmatrix}$$

$$T_2 = \begin{pmatrix} \frac{2-x-y}{2} & 0 & \frac{x}{2} & \frac{y}{2} \\ 1-y & 0 & 0 & y \\ \frac{2(1-y)+(1-x)(2-\phi)}{4} & \frac{\phi}{8} & \frac{x(2-\phi)}{4} & \frac{y}{2} \\ \frac{2n(1-y)+(n-1)(1-x)(1-\phi)}{4n} & \frac{\phi(n-1)+2}{8n} & \frac{(n-1)(1+x(1-\phi))}{4n} & \frac{y}{2} \end{pmatrix}$$

$$T_3 = \begin{pmatrix} \frac{4-2yA-x(d_m+d_f)}{4} & 0 & \frac{x(d_m+d_f)}{4} & \frac{yA}{2} \\ (1-y)A & 0 & 0 & 1-(1-y)A \\ \frac{2(1-y)A+B}{4} & \frac{\phi}{8} & \frac{2-\phi-B}{4} & \frac{1-(1-y)A}{2} \\ \frac{2n(1-y)A+(n-1)B}{4n} & \frac{\phi(n-1)+2}{8n} & \frac{(n-1)(2-\phi-B)}{4n} & \frac{1-(1-y)A}{2} \end{pmatrix} \tag{12}$$

Columna vector constante:

Se define de la siguiente manera:

$$C = \begin{pmatrix} 0 \\ 0 \\ \frac{\phi}{8} \\ \frac{\phi * (n-1) + 2}{8 * n} \end{pmatrix} \tag{13}$$

Estadísticos F:

Se calculan con los estadísticos F para el tiempo *t*, extrayendo los parámetros de coancestralidad de la columna vector S_t, así:

$$F_{LS} = \frac{\theta - \alpha}{1 - \alpha};\ F_{IL} = \frac{F - \theta}{1 - \theta};\ F_{IS} = \frac{F - \alpha}{1 - \alpha} \tag{14}$$

Se almacenan los estadísticos F obtenidos para el tiempo $t = 0$, que son iguales a cero en el caso por defecto, o se calculan a partir de las frecuencias genotípicas en caso de poseer información de entrada de un formato de Genepop. Se debe repetir el procedimiento para cada instante de tiempo *t*. Para calcular el nuevo vector columna S_{t+1} se debe realizar la siguiente operación:

$$S_{t+1} = T * S_t + C \tag{15}$$

Del cuál calculamos los nuevos estadísticos F. Cada nuevo conjunto de datos calculado será almacenado para su posterior graficación. Este proceso iterativo se realiza para la matriz de transición T_1, T_2 o T_3, dependiendo del caso escogido por el usuario, por un número dado de generaciones especificadas por el usuario. Una vez calculados todos los estadísticos F se grafican los resultados para cada caso de estudio (matrices T_1, T_2 y T_3) de forma paralela, para su comparación.

Diagrama de flujo

El diagrama de flujo fue efectuado con la ayuda del programa SFC (Structured Flowchart Creator) Versión 2.3, Copyright (C) 2000, creado por Tia Watts, Ph. D., Sonoma State University. Las Figuras 1 - 7 muestran el esquema general del programa, la rutina principal y sus 5 subrutinas.

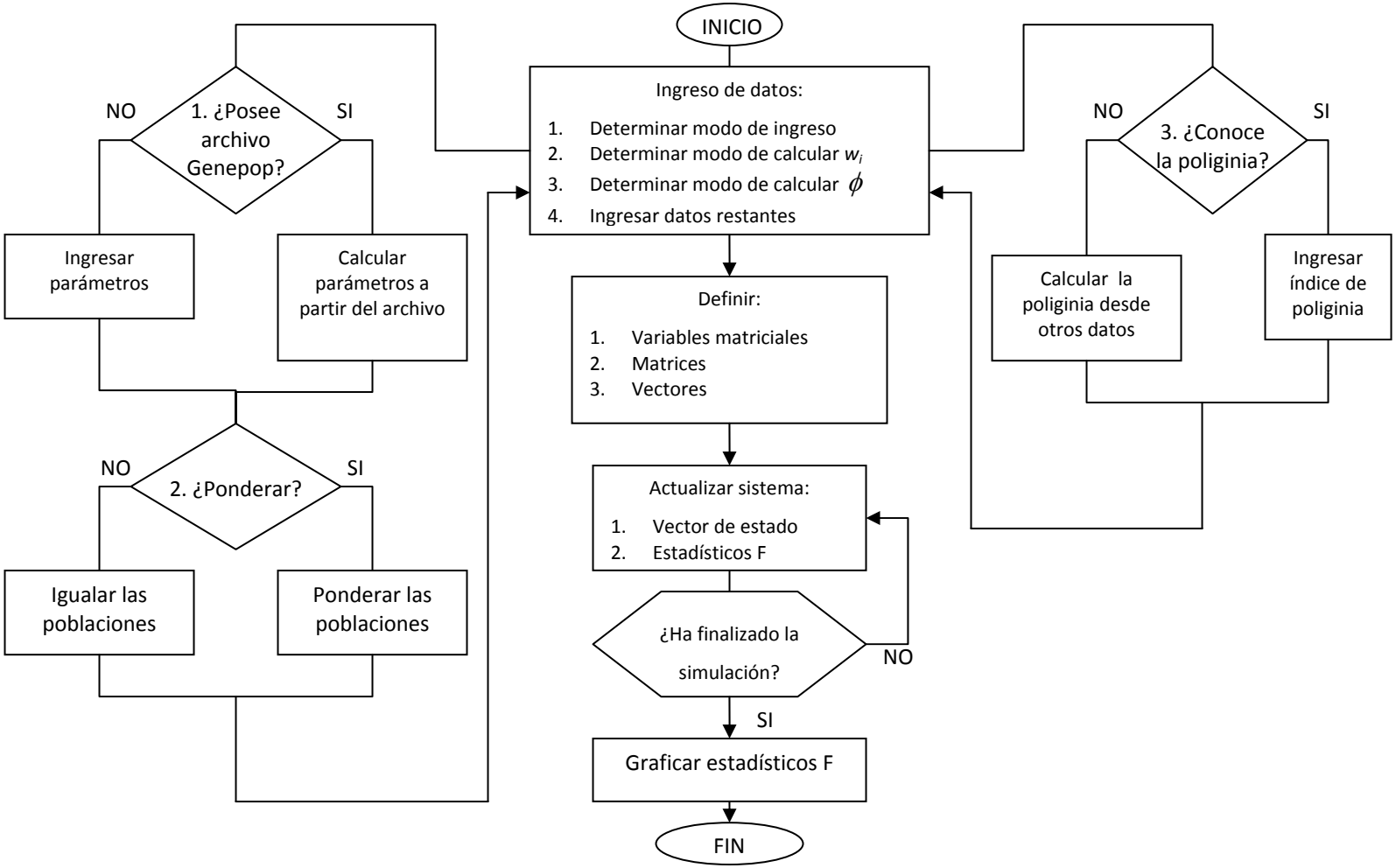


Figura 1 Diagrama de flujo esquemático de FstatSim

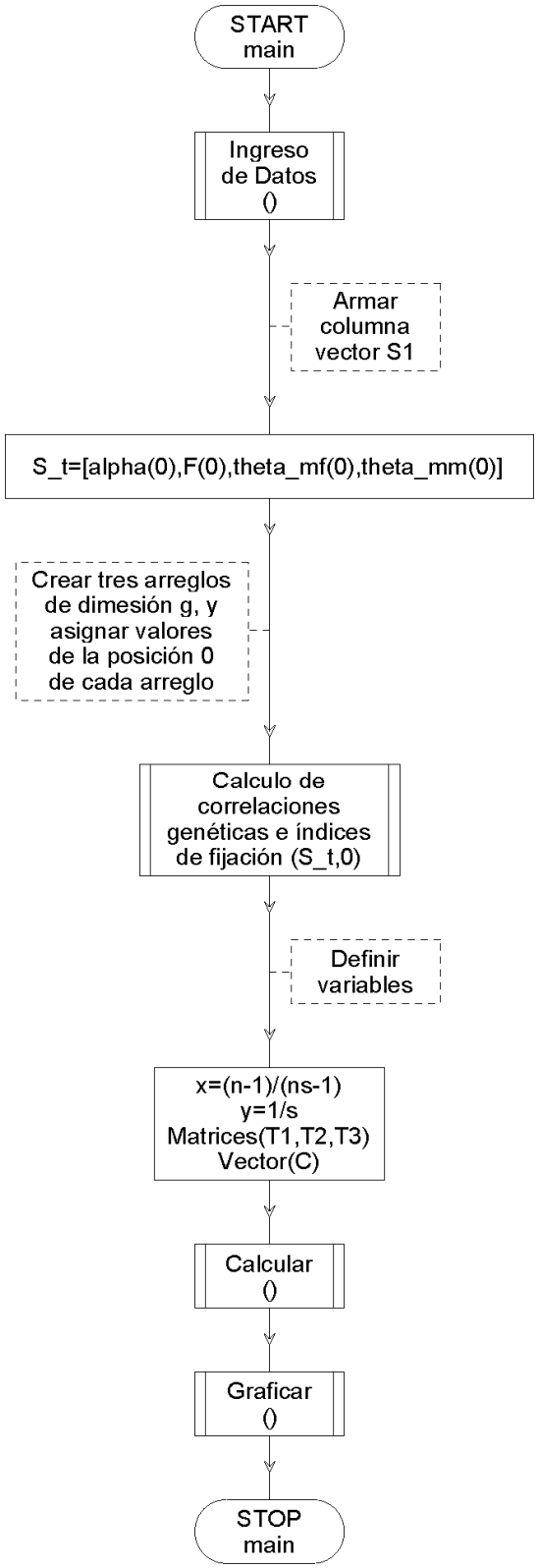


Figura 2 Diagrama de flujo: Rutina principal de FstatSim.

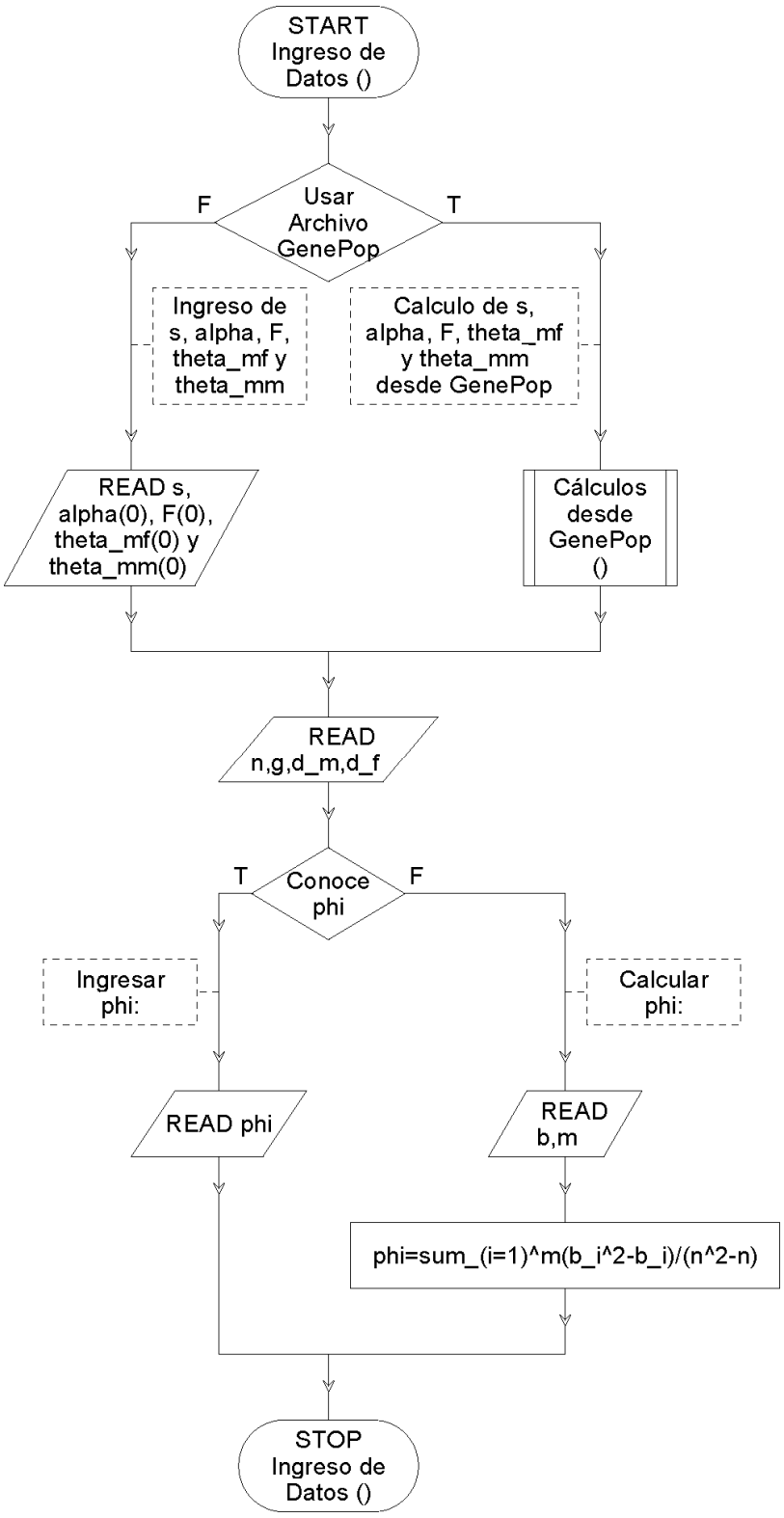


Figura 3 Diagrama de flujo: Subrutina 'Ingreso de Datos' de FstatSim.

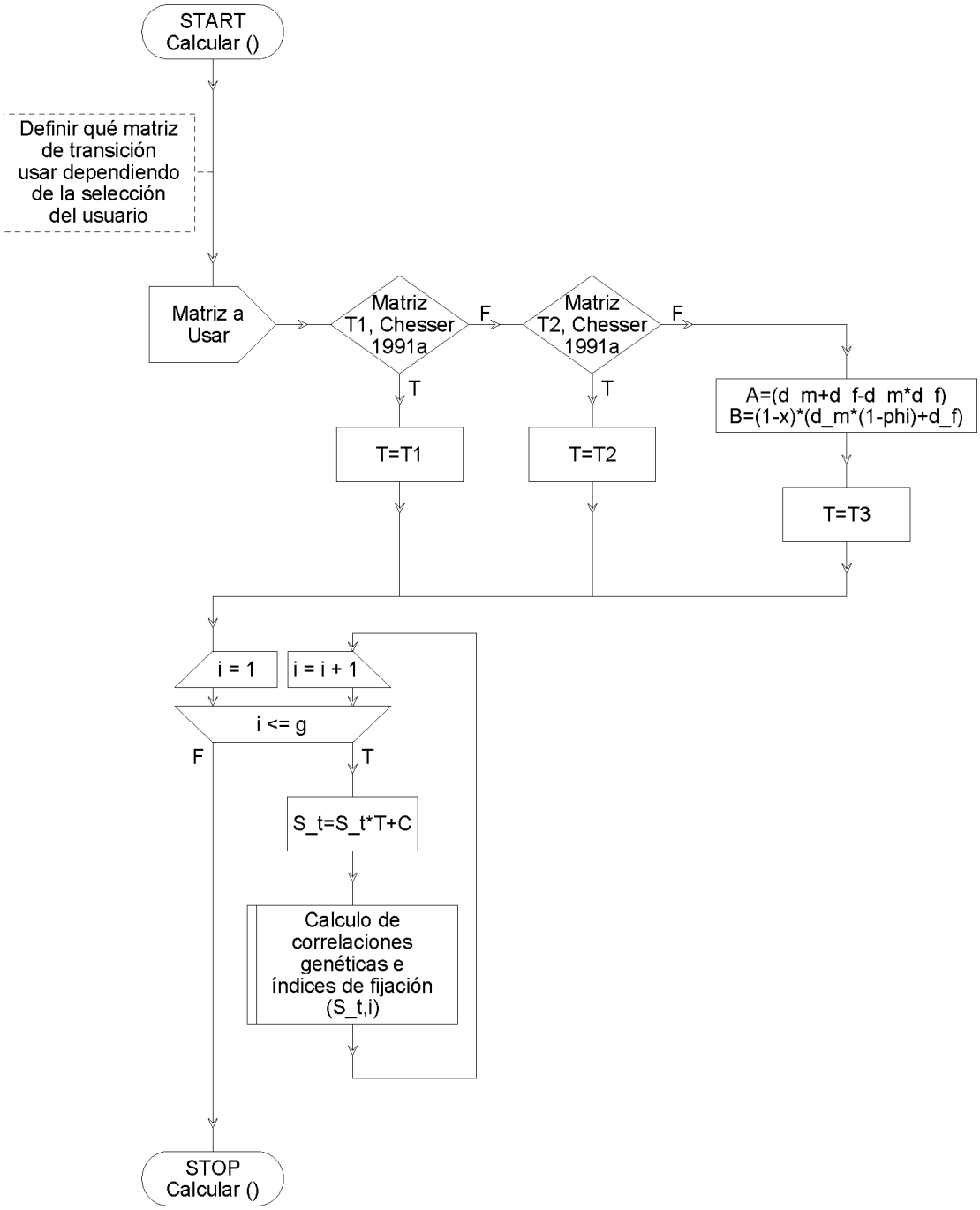


Figura 4 Diagrama de flujo: Subrutina 'Calcular' de FstatSim.

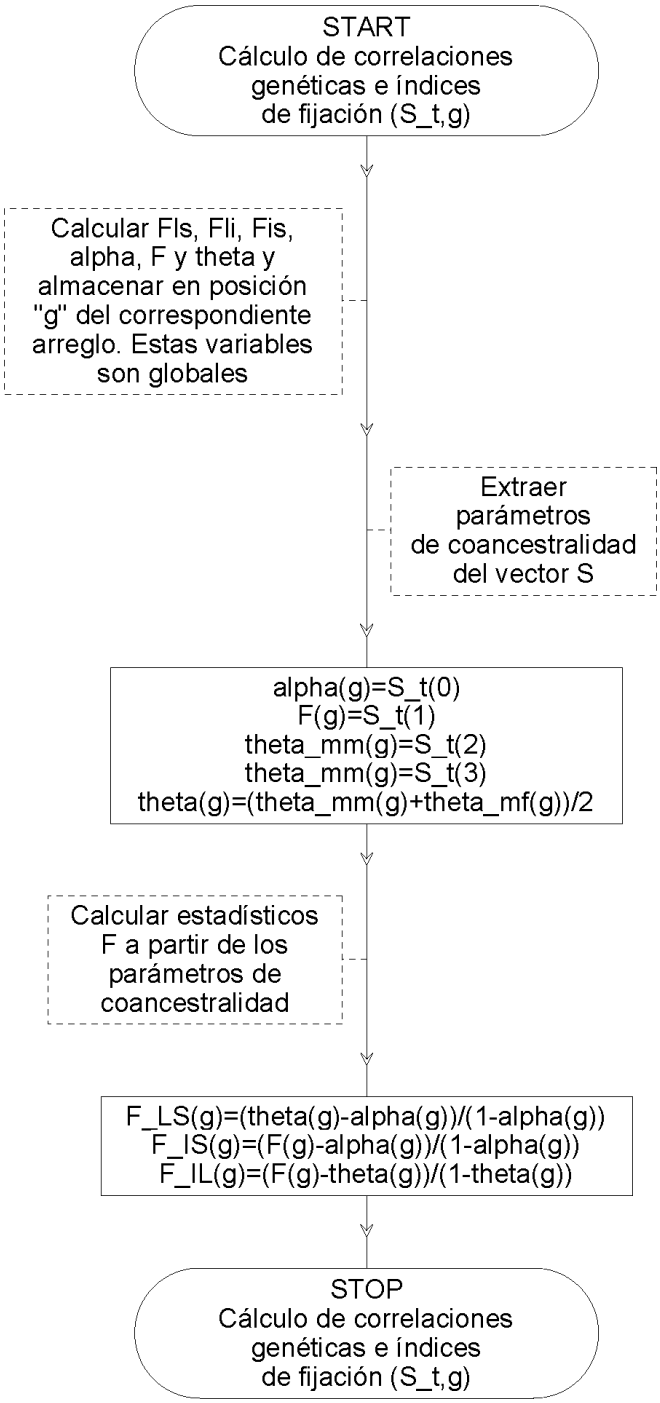


Figura 5 Diagrama de flujo: Subrutina 'Cálculo de correlaciones genéticas e índices de fijación' de FstatSim.

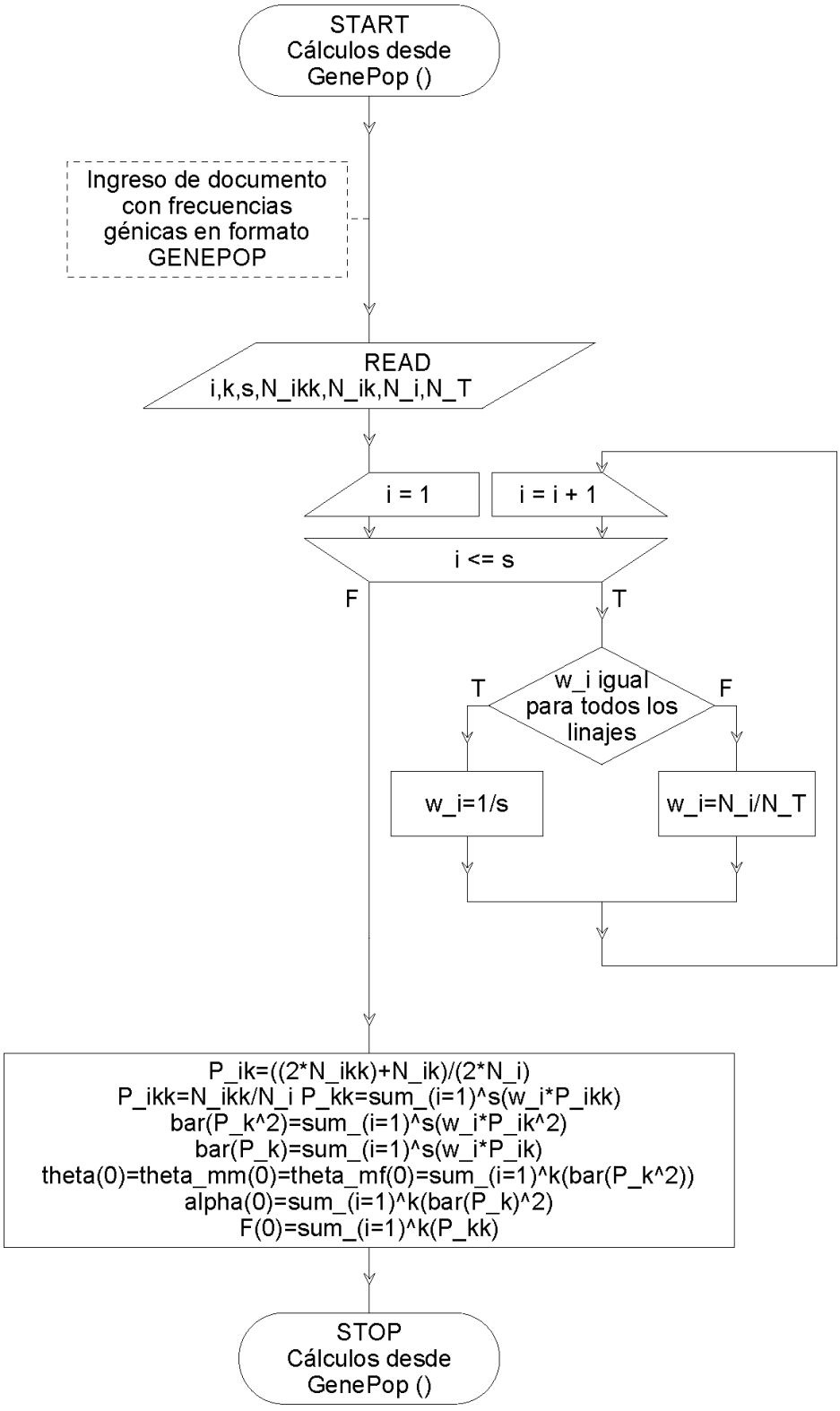


Figura 6 Diagrama de flujo: Subrutina 'Cálculos desde GenePop' de FstatSim.

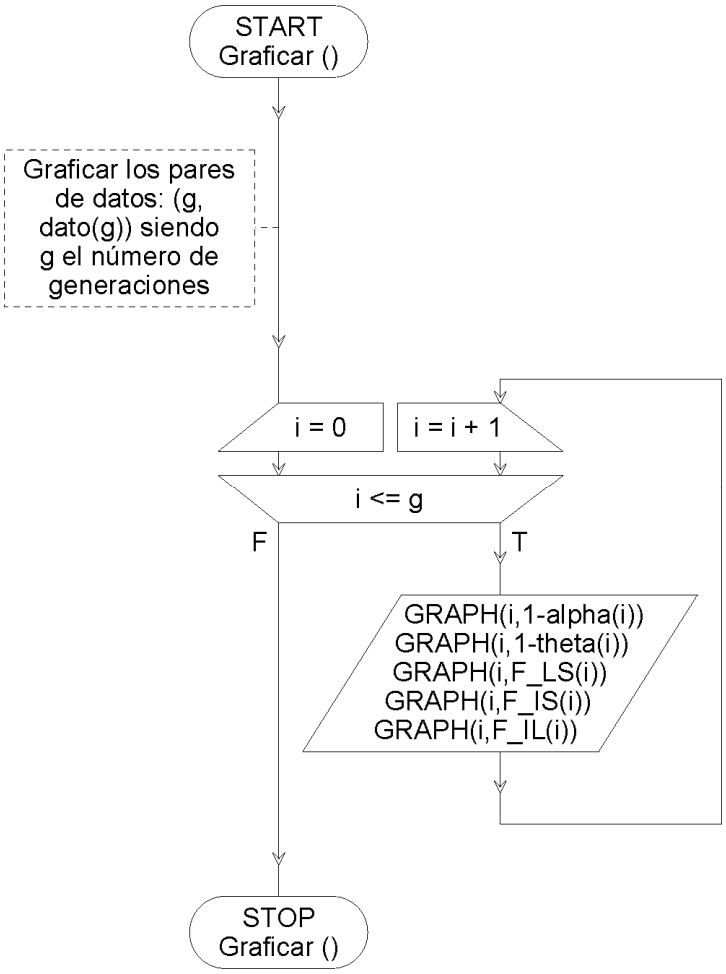


Figura 7 Diagrama de flujo: Subrutina 'Graficar' de FstatSim.

5.2.3 Programación

Todo el programa fue codificado usando Visual Studio 2005 Express Edition en el lenguaje Visual Basic .NET. El código fuente resultante de esta fase se presenta en el software.

5.2.4 Pruebas

Las actividades de esta fase son las correspondientes al **cuarto objetivo**. Durante la codificación se efectuaron pruebas de validación de los datos, es decir, comprobaciones para asegurar que los datos introducidos, ya fuera manualmente o al

cargar archivos válidos de Genepop, mantuvieran una validez desde el punto de vista matemático y biológico.

Se realizó además una comparación gráfica de los resultados obtenidos con los esperados. Chesser (1991a, b) presentó gráficos que demostraban los resultados de las computaciones descritas, para comparar los efectos de las diferentes variables sobre los índices de fijación calculados. Se obtuvo una conformidad satisfactoria entre estos gráficos y los producidos por FStatSim para los mismos datos iniciales. Los gráficos son presentados y discutidos en la sección de resultados y discusión.

El comportamiento del sistema en cuanto a rendimiento se evaluó invocando el manejador de tareas de Windows durante la ejecución del programa. Se observó que éste usa un recurso de cerca de 20 MB de RAM, lo cuál es una cantidad media. En cuanto a velocidad, usabilidad y apariencia, el programa fue sometido a constantes pruebas, por parte de varios usuarios que se tomaron el trabajo de leer la teoría implicada para comprender el propósito del programa. Se realizaron pruebas de concepto, en las que diferentes versiones sucesivas del programa fueron enviadas a distintos usuarios (estudiantes de biología familiarizados con software de manejo estadístico) para tomar nota de los comentarios cualitativos suscitados. Las evaluaciones cualitativas de los usuarios en cuanto a los aspectos de usabilidad y apariencia, mejoraron con cada nueva versión, mientras que la velocidad nunca fue motivo de preocupación (la realización de la simulación ocurre a gran velocidad y de manera robusta, incluso para valores muy grandes de número de generaciones). La tabla que muestra este progreso se presenta en la sección de resultados.

El único error (esporádico) conocido es que el programa se traba al tratar de volver a cargar un archivo de Genepop que ya está cargado (por ejemplo, cuando se quiere volver a correr la simulación con diferentes parámetros de entrada).

5.2.5 Documentación

Documentación Técnica e Interna

La documentación técnica y la interna (comentarios en el código fuente) fueron realizadas durante y después de la fase de programación.

Las fases de Análisis y Diseño complementan esta documentación pero no se incluyen para evitar duplicar contenidos en este documento.

Documentación del Usuario

Para la documentación del usuario se usó la modalidad tutorial, junto con la modalidad de listas, y se incluyeron comentarios técnicos acerca de la realización del software. A continuación una adaptación del documento que se incluye como PDF en la distribución del programa.

Ingreso de datos

El programa valida los datos, alertando cuando no están dentro de los rangos apropiados. Debe tener en consideración el separador de decimales (punto o coma).

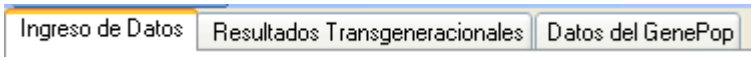


Figura 8 Pestaña principal

Para facilitar su entendimiento, se incluyen las siguientes tablas explicativas de las variables de entrada, su descripción, el tipo de variable y sus rangos permitidos. (Tabla 1 y Tabla 2)

Las correlaciones genéticas, o parámetros de coancestralidad (α, F, θ_{mm} y θ_{mf}) pueden ingtresarse manualmente o calcularse a partir de un archivo de formato Genepop.

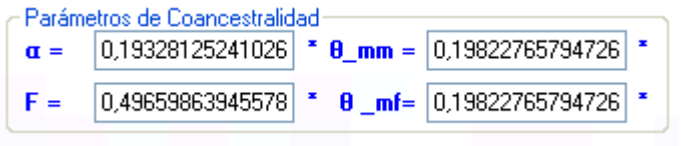


Figura 9 Ingreso de parámetros de coancestralidad

El ingreso de las tasas de dispersión (d_m y d_f) se divide en tres casos, uno para cada matriz de transición usada. Los primeros dos casos asumen valores fijos de d_m y d_f, y el tercer caso permite el ingreso de valores distintos para d_m y d_f.

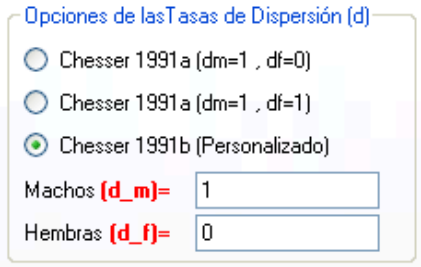


Figura 10 Opciones de tasas de dispersión

Digite n, s, m, b y g ingresando los valores deseados. Si planea usar un archivo Genepop tenga en cuenta que de ahí también se extrae el número de linajes "s", por lo tanto no debe modificarlo (aplica también para los parámetros de coancestralidad).

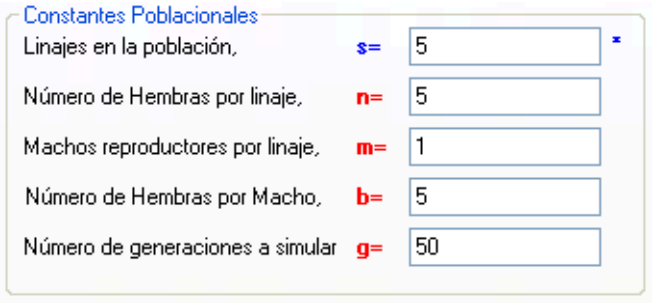


Figura 11 Ingreso de constantes poblacionales

Puede seleccionar entre calcular ϕ a partir de los valores ingresados de n, m y b, o ingresar ϕ directamente.

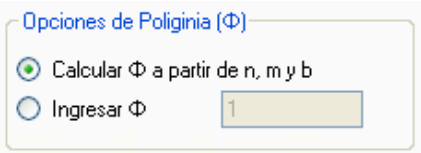


Figura 12 Opciones de poliginia (phi)

Cargar datos desde un Archivo de Genepop

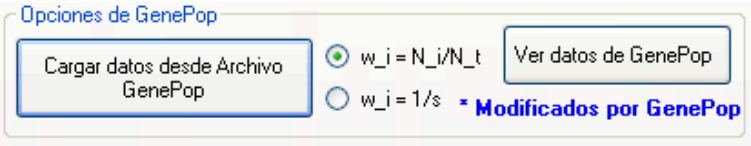


Figura 13 Opciones de Genepop

El ingreso de correlaciones genéticas (α, F, θ_{mm} y θ_{mf}) puede hacerse introduciendo los valores en cajas de texto, o usando las opciones de Genepop. Estas permiten cargar los datos de un archivo de Genepop previamente creado.

Figura 14 Botón para cargar un archivo Genepop

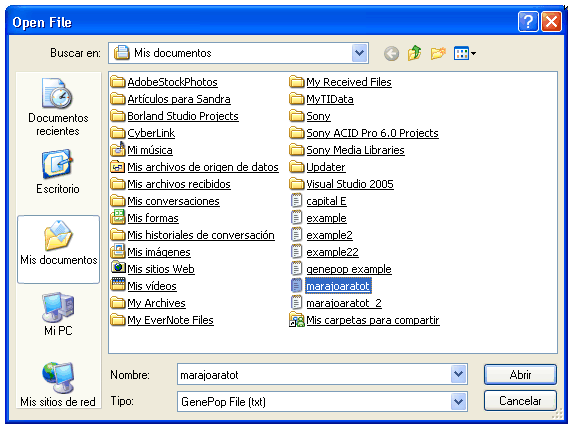


Figura 15 Cuadro de diálogo

En el momento de cargar el archivo de Genepop, existe la opción, mediante botones radiales, de calcular w_i mediante dos métodos: uno asumiendo una ponderación equitativa entre linajes, y otro que atribuye ponderaciones distintas dependiendo del número de individuos en cada linaje.

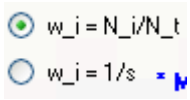


Figura 16 Opciones de cálculo de factor de ponderamiento poblacional

Calcular y graficar

Una vez ingresados todos los datos presione el botón "Simular" para calcular los índices de fijación y graficarlos.

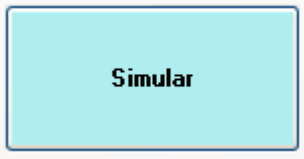


Figura 17 Botón para realizar la simulación

Gráfica

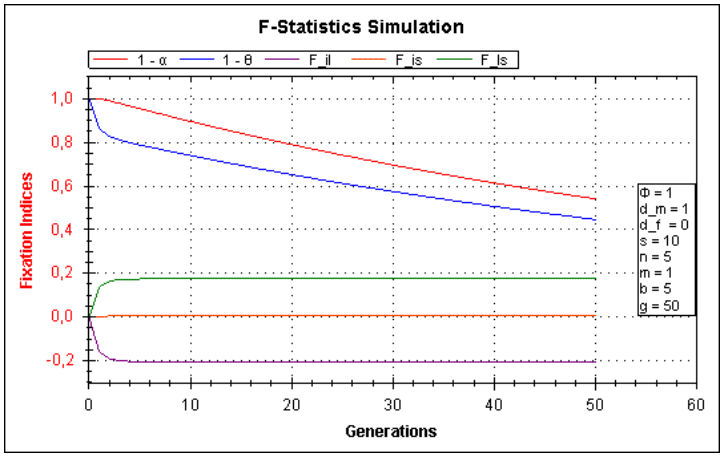


Figura 18 Gráfica de la simulación predeterminada en el programa

Una vez simulados los datos ingresados el programa genera una gráfica para facilitar la visualización de la simulación y sus datos numéricos. Esta puede ser manipulada para su comodidad. Tenga en cuenta los siguientes comandos:

Zoom: mouse izq. y arrastrar Mover: mouse med. y arrastrar Menú Contextual: mouse der. Info del punto: mouse sobre

Opciones de estilo de la gráfica

Figura 19 Opciones de estilo de la gráfica

Puede seleccionar las opciones de visualización de la gráfica según lo desee. Estas opciones se dan para facilitar la legibilidad de la gráfica y su uso en otros documentos.

Opciones adicionales de la gráfica

Aparte de estas opciones, puede acceder a otros recursos haciendo clic derecho sobre la gráfica:

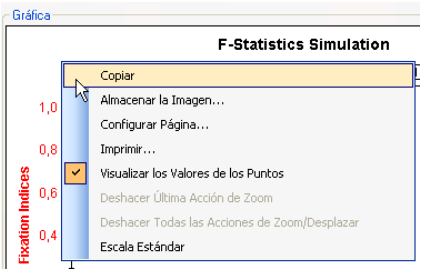


Figura 20 Opciones adicionales de la gráfica

Copiar la gráfica

Esta opción le permite “Copiar” al clipboard la imagen para pegarla en documentos u otras aplicaciones. **Nota:** Generalmente es mejor guardar la imagen y después insertarla en vez de copiarla y pegarla. Éste último método requiere más memoria.

Guardar la grafica

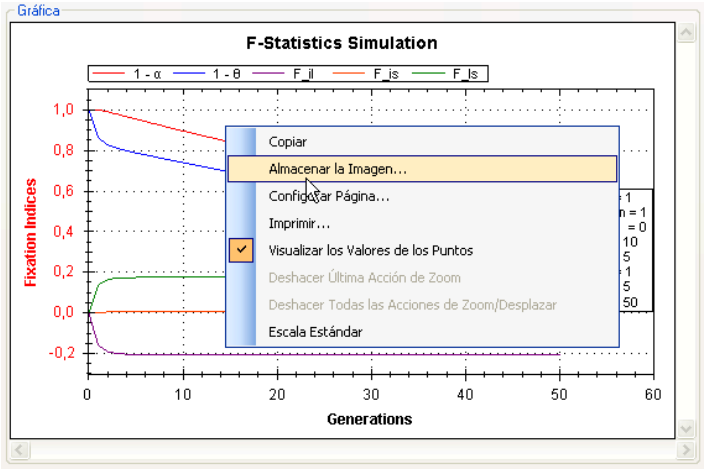


Figura 21 Opción para almacenar la imagen

El programa también le permite guardar la imagen como un archivo de imagen para facilitar su manipulación y reproducción en otras aplicaciones. Al seleccionar la opción “Almacenar la Imagen…” aparecerá un cuadro de diálogo para que usted seleccione el nombre, destino y formato del archivo.

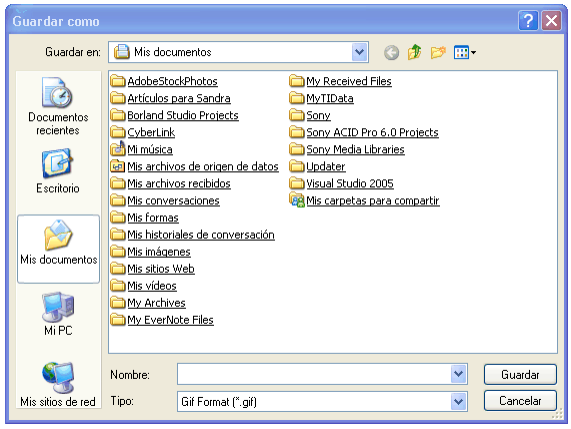


Figura 22 Cuadro de diálogo

Los formatos disponibles son: (*.emf), (*.png), (*.gif), (*.jpg), (*.tif), o (*.bmp). Cada formato cuenta con distintas características que no serán explicadas en este manual. Seleccione la de mayor conveniencia:

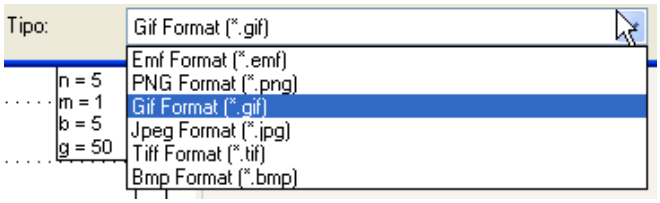


Figura 23 Formatos de imagen disponibles

Opciones de Impresión

Con las opciones "Configurar Página…" e "Imprimir…" puede imprimir la gráfica. El documento impreso sólo contará con la imagen y los datos que ella contenga.

Opciones adicionales de visualización

Con las últimas tres opciones puede modificar cambios de Zoom que haya realizado, así como modificar la escala en la que se muestran los datos.

Tablas

Una vez realizada la simulación también se obtiene una tabla de resultados en donde aparecen los índices de fijación para cada generación:

Ingreso de Datos | Resultados Transgeneracionales | Datos del GenePop

Estadísticos-F: Resultados Transgeneracionales

Gen	F_il	F_is	F_ls	1 - α	1 - θ
0	0	0	0	1	1
1	-0,159420289855072	0	0,1375	1	0,8625
2	-0,191232048374906	0,00499922692367159	0,164729680977168	0,989948979591837	0,826875
3	-0,200949610037835	0,00560804765749449	0,171995274380265	0,978189816743024	0,809945790816327
4	-0,203383442145441	0,00577971130276867	0,173812557263797	0,966089759050226	0,79817122748334
5	-0,204005001308648	0,00582139655083462	0,174273692909431	0,954021172707761	0,787760379826194
6	-0,204162306686038	0,00583200763327518	0,174390373418378	0,942073717291436	0,777785129945343
7	-0,204202149464875	0,00583468917745411	0,174419916735463	0,930268480407488	0,768011129513188

Figura 24 Tabla de resultados transgeneracionales

Es posible llegar a la tabla usando el botón “Ver Resultados Transgeneracionales” que aparece en la pestaña “Ingreso de Datos”, debajo de la gráfica, o sencillamente haciendo clic en la pestaña “Resultados Transgeneracionales”:

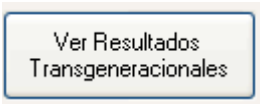


Figura 25 Botón para ver los resultados transgeneracionales

Tablas de Datos del Archivo GenPop

Si uso un archivo de Genepop para calcular los parámetros iniciales de coancestralidad y el número de linajes, en la tercera pestaña: “Datos del Genepop” aparecerán los datos extraídos del mismo

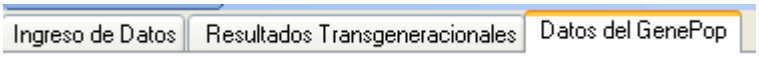


Figura 26 Pestaña de Datos del archivo de Genepop cargado

Las tres tablas allí presentes serán llenadas con los datos del Archivo Genepop. Esta información es útil para entender la forma en que se realizan los cálculos, partiendo de que el programa usa el método del artículo de Nei (1977).

Los valores calculados son las frecuencias genotípicas, las frecuencias alélicas, homocigóticas y heterocigóticas por población, las frecuencias homocigóticas y heterocigóticas por alelo, los loci etiquetados, el número de individuos y valores de w_i por población, los parámetros de coancestralidad momentáneos, y los estadísticos F momentáneos, junto con una comprobación de la regla de Wright a partir de ellos. Las tablas permiten ser copiadas y pegadas como texto normal en tablas de Excel.

Exportar Tablas de Datos a Excel

El programa le permite exportar todas las tablas obtenidas a un formato legible por Excel (XLS). Esto le da la opción al usuario de manipular los datos para reproducir la gráfica con otras configuraciones visuales, incluir la información en un reporte o artículo, etc.

Debajo de cada tabla, en el lado derecho, encontrará un botón marcado como "Exportar a Excel".

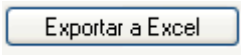


Figura 27 Botón de exportación a Excel

Una vez presionado, un cuadro de diálogo aparecerá. Allí podrá seleccionar el nombre para el archivo y su destino.

Abrir y Guardar Conjuntos de Datos

El programa guarda automáticamente los últimos datos usados cada vez que se cierra la aplicación. Aparte de esto, el usuario también tiene la opción de almacenar los datos ingresados para usarlos en otro momento. El formato de los archivos generados por el programa es (*.fsm)

Abrir

Al seleccionar la opción Abrir, ya sea desde el menú, el icono o la combinación de teclas CONTROL+A, aparecerá un cuadro de diálogo que le permite buscar un archivo (*.fsm) previamente almacenado para cargar una sesión completa y reproducir la simulación.

Guardar

Al seleccionar la opción Guardar, ya sea desde el menú, el icono o la combinación de teclas CONTROL+G, aparecerá un cuadro de diálogo que le permite seleccionar el nombre para el archivo y su destino. De esta forma podrá reproducir una simulación en cualquier otro momento.

Ayuda

FStatSim fue creado usando un componente desarrollado por terceras partes, esto significa que pueden existir errores sobre los cuales no se tenga dominio (en el componente de graficación).

Si encuentra algún error, por favor contribuya con el proyecto enviando al autor una descripción detallada del fallo para que este pueda ser corregido y otros puedan beneficiarse de nuevas y mejores versiones.

5.2.6 Implementación

El programa ocupa cerca de 500 Kb, de manera que se puede cargar fácilmente en un dispositivo de memoria USB y no requiere instalación previa. Sin embargo, requiere que el sistema operativo posea una versión reciente de .NET Framework instalada. Esta se consigue gratuitamente por Internet.

5.3 Recolección de la Información

Al tratarse de un desarrollo tecnológico, este trabajo carece de tratamiento recolección de información en campo. La información recolectada para este trabajo fue solo bibliográfica.

5.4 Análisis de Información

Al tratarse de un desarrollo tecnológico, este trabajo carece de tratamiento estadístico de los datos.

6. Resultados y Discusión

El resultado del presente trabajo de grado es el programa FstatSim validado, en versiones opcionales en inglés o en español, con su respectiva documentación técnica (Anexo 3) y disponible en línea en <https://sourceforge.net.projects/fstatsim> (todo bajo licencia GNU). El programa parte de los parámetros de coancestralidad (F, θ y α), obtenidos como datos iniciales, o a partir de las frecuencias genotípicas de un formato de Genepop, para armar con ellos un vector que será multiplicado con una de tres matrices distintas de transición (T_1, T_2 y T_3), dependiendo del caso biológico seleccionado (hembras totalmente filopátricas y machos totalmente migrantes; ambos sexos totalmente migrantes; o cualquier combinación de valores de dispersión para ambos sexos, respectivamente, junto con otras constantes poblacionales), y sumado con un vector constante (C). Este proceso, repetido un número de veces equivalente al número de generaciones solicitadas por el usuario, permite obtener un nuevo vector de parámetros de coancestralidad, estimados para el número de generaciones futuras establecido. A partir de los parámetros de coancestralidad, el programa evalúa para cada generación los coeficientes F para los niveles individuo-linaje-subpoblación (F_{LS}, F_{IL} y F_{IS}), donde la subpoblación se considera como la población total, y traza las curvas de cinco índices ($1 - \alpha$, $1 - \theta$, F_{LS}, F_{IL} y F_{IS}) en una gráfica que muestra la transición continua de estos valores a lo largo de las generaciones. Al permitir escoger diferentes valores de entrada tanto para los parámetros de coancestralidad como para otras constantes poblacionales que describen la estrategia reproductiva y la tasa de dispersión sexual, se puede observar la influencia relativa de estos diferentes valores en la transición de los índices de fijación.

Las Figuras 9, 10 y 11 muestran, respectivamente, las tres pantallas que presenta el programa: ingreso de datos, ejecución de simulación y graficación de resultados; la de presentación de datos transgeneracionales de los parámetros simulados; y la de presentación de algunos datos obtenidos del archivo Genepop para su comprobación o registro si se desea.

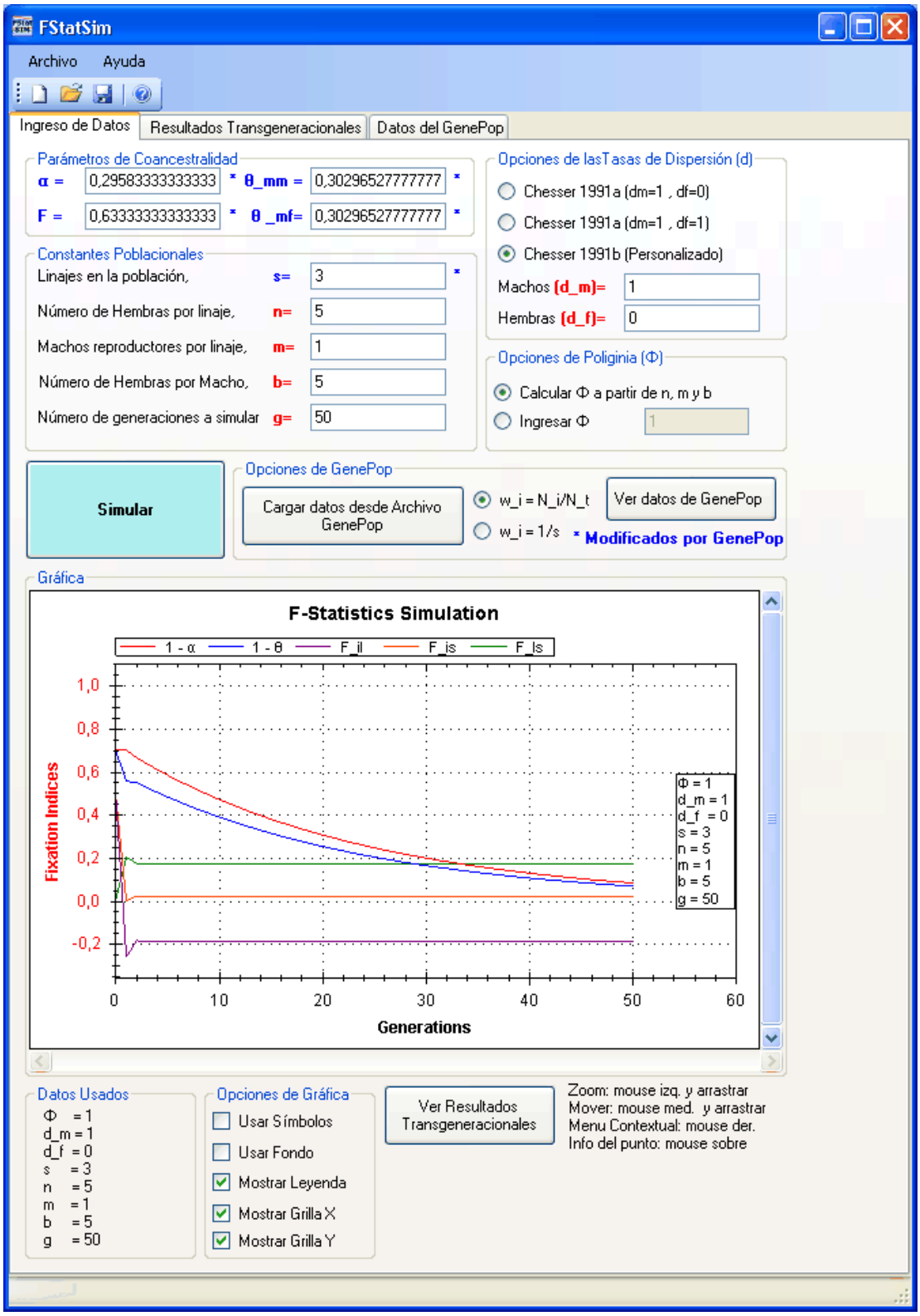
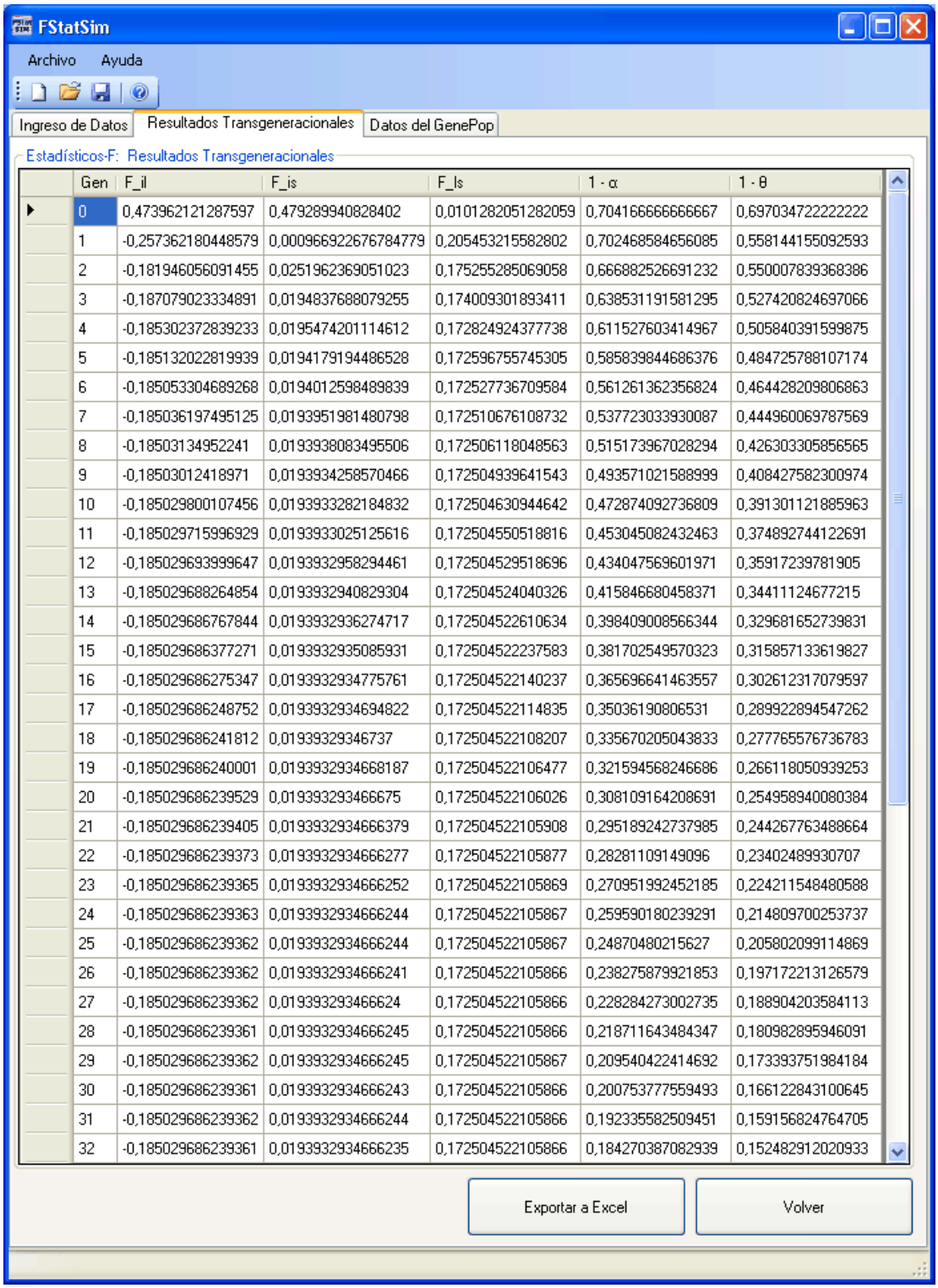


Figura 28 Ingreso de datos, ejecución de simulación y graficación de resultados



Gen	F_il	F_is	F_ls	1 - α	1 - θ
0	0,473962121287597	0,479289940828402	0,0101282051282059	0,704166666666667	0,697034722222222
1	-0,257362180448579	0,000966922676784779	0,205453215582802	0,702468584656085	0,558144155092593
2	-0,181946056091455	0,0251962369051023	0,175255285069058	0,666882526691232	0,550007839368386
3	-0,187079023334891	0,0194837688079255	0,174009301893411	0,638531191581295	0,527420824697066
4	-0,185302372839233	0,0195474201114612	0,172824924377738	0,611527603414967	0,505840391599875
5	-0,185132022819939	0,0194179194486528	0,172596755745305	0,585839844686376	0,484725788107174
6	-0,185053304689268	0,0194012598489839	0,172527736709584	0,561261362356824	0,464428209806863
7	-0,185036197495125	0,0193951981480798	0,172510676108732	0,537723033930087	0,444960069787569
8	-0,18503134952241	0,0193938083495506	0,172506118048563	0,515173967028294	0,426303305856565
9	-0,18503012418971	0,0193934258570466	0,172504939641543	0,493571021588999	0,408427582300974
10	-0,185029800107456	0,0193933282184832	0,172504630944642	0,472874092736809	0,391301121885963
11	-0,185029715996929	0,0193933025125616	0,172504550518816	0,453045082432463	0,374892744122691
12	-0,185029693999647	0,0193932958294461	0,172504529518696	0,434047569601971	0,35917239781905
13	-0,185029688264854	0,0193932940829304	0,172504524040326	0,415846680458371	0,34411124677215
14	-0,185029686767844	0,0193932936274717	0,172504522610634	0,398409008566344	0,329681652739831
15	-0,185029686377271	0,0193932935085931	0,172504522237583	0,381702549570323	0,315857133619827
16	-0,185029686275347	0,0193932934775761	0,172504522140237	0,365696641463557	0,302612317079597
17	-0,185029686248752	0,0193932934694822	0,172504522114835	0,35036190806531	0,289922894547262
18	-0,185029686241812	0,01939329346737	0,172504522108207	0,335670205043833	0,277765576736783
19	-0,185029686240001	0,0193932934668187	0,172504522106477	0,321594568246686	0,266118050939253
20	-0,185029686239529	0,019393293466675	0,172504522106026	0,308109164208691	0,254958940080384
21	-0,185029686239405	0,0193932934666379	0,172504522105908	0,295189242737985	0,244267763488664
22	-0,185029686239373	0,0193932934666277	0,172504522105877	0,28281109149096	0,23402489930707
23	-0,185029686239365	0,0193932934666252	0,172504522105869	0,270951992452185	0,224211548480588
24	-0,185029686239363	0,0193932934666244	0,172504522105867	0,259590180239291	0,214809700253737
25	-0,185029686239362	0,0193932934666244	0,172504522105867	0,24870480215627	0,205802099114869
26	-0,185029686239362	0,0193932934666241	0,172504522105866	0,238275879921853	0,197172213126579
27	-0,185029686239362	0,019393293466624	0,172504522105866	0,228284273002735	0,188904203584113
28	-0,185029686239361	0,0193932934666245	0,172504522105866	0,218711643484347	0,180982895946091
29	-0,185029686239362	0,0193932934666245	0,172504522105867	0,209540422414692	0,173393751984184
30	-0,185029686239361	0,0193932934666243	0,172504522105866	0,200753777559493	0,166122843100645
31	-0,185029686239362	0,0193932934666244	0,172504522105866	0,192335582509451	0,159156824764705
32	-0,185029686239361	0,0193932934666235	0,172504522105866	0,184270387082939	0,152482912020933

Figura 29 Presentación de datos transgeneracionales de los parámetros simulados.

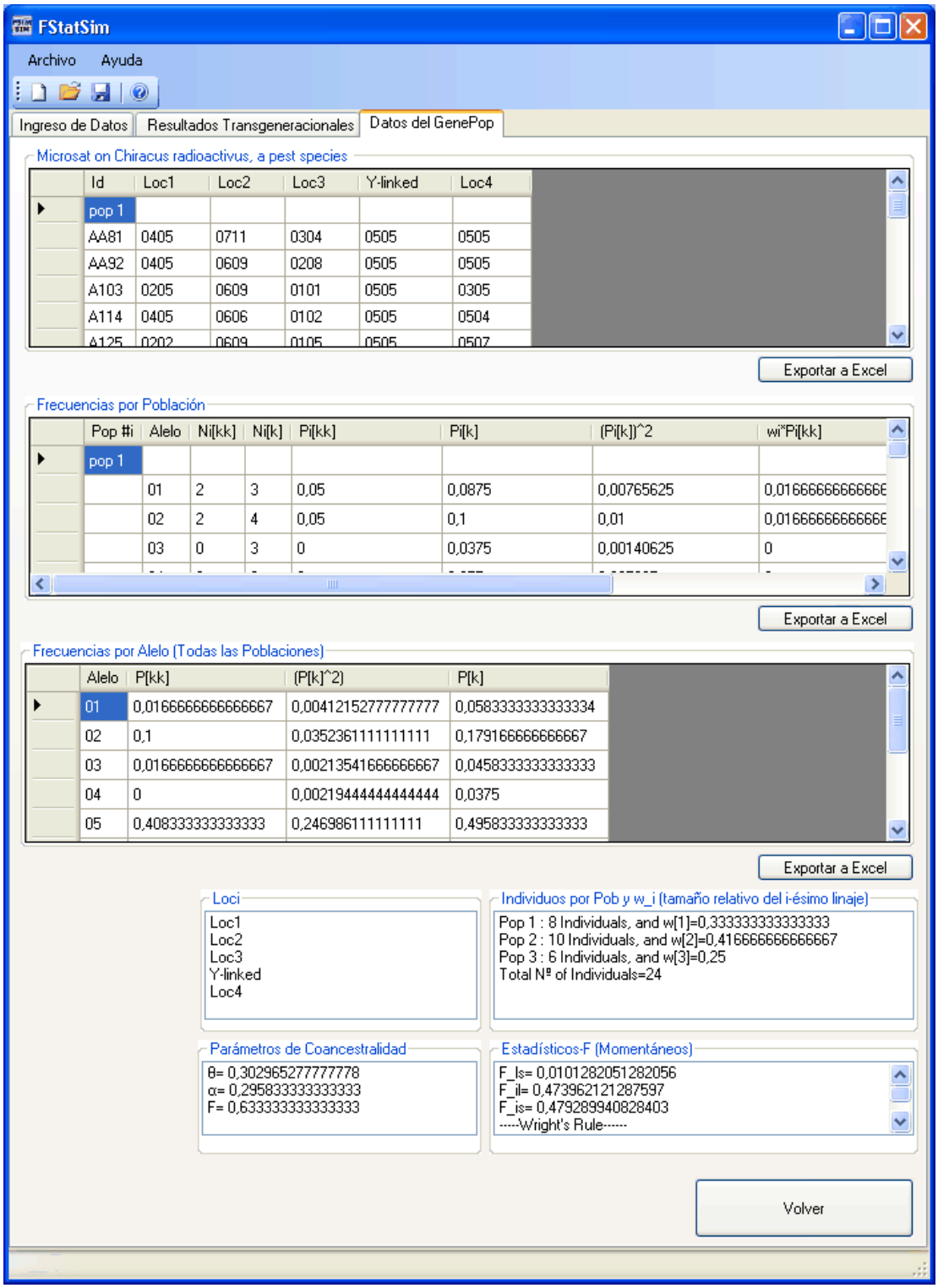
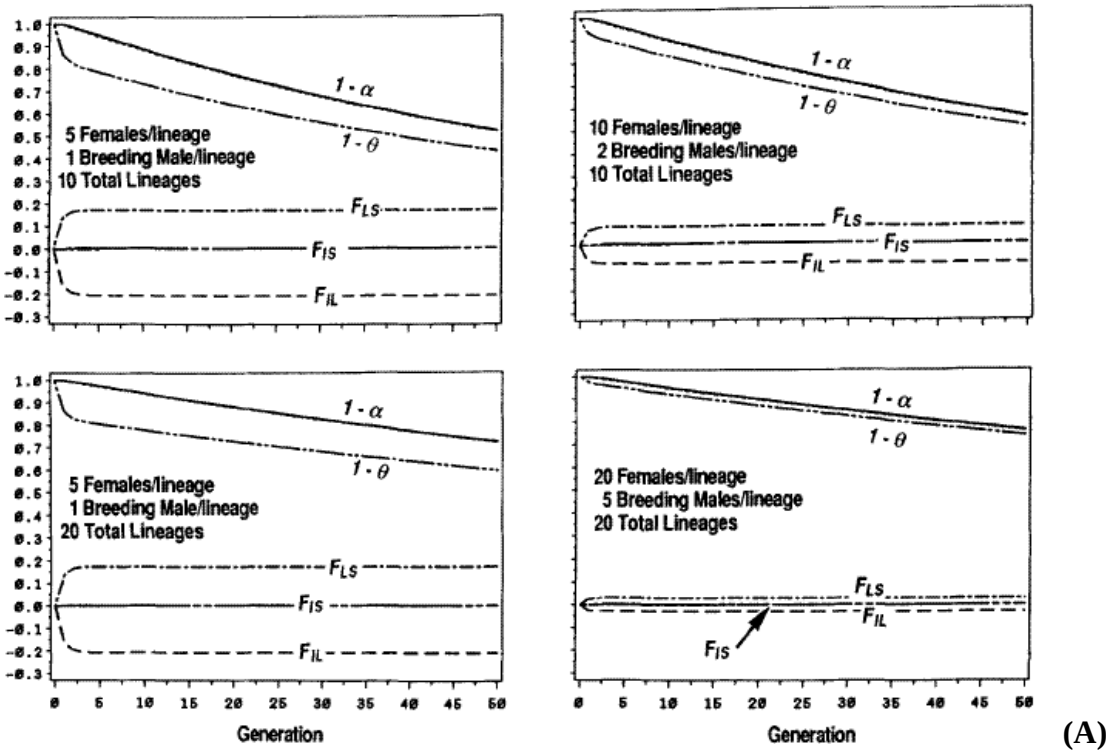


Figura 30 Presentación de algunos datos obtenidos del archivo Genepop.

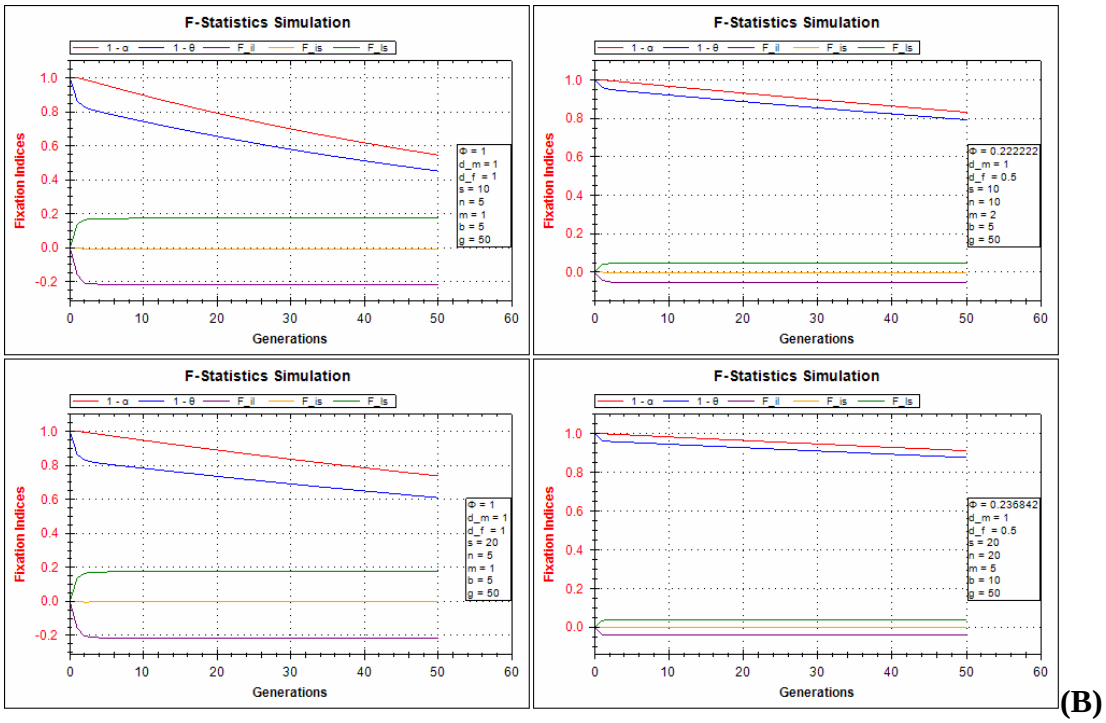
COMPARACIÓN DE RESULTADOS OBTENIDOS Y ESPERADOS:

La Figura 31 y la Figura 32 muestran los cambios en índices de fijación y correlaciones genéticas a través de las generaciones como resultado de diferentes escenarios reproductivos, tasas migratorias y no-independencia de parejas de machos (ϕ): $1 - \alpha$ es la proporción de la variancia genética original que permanece en la población;1- θ es la que permanece dentro de los linajes; F_{LS} es la proporción de diversidad genética en cualquier generación, que se encuentra entre los linajes, F_{IS} es la correlación de genes dentro de individuos (endogamia) relativa a la que se halla dentro de la (sub)población, y F_{IL} es la correlación de genes dentro de individuos relativa a los disponibles dentro de los linajes.

Las figuras muestra además una comparación lado a lado de los gráficos obtenidos por Chesser (Figura 10 (A) y Figura 11(A))y los obtenidos por FStatSim (Figura 10 (B) y Figura 11(B)), usando los mismos parámetros (hasta donde son especificados en los gráficos presentados por Chesser). Se observa que las gráficas obtenidas por FstatSim son más informativas (al momento de ser generadas se pueden deslizar y redimensionar, además de que los datos precisos de cualquier punto de cualquier curva se pueden consultar llevando el puntero sobre ese punto, con el mouse). Se observa además que aunque las gráficas comparadas de Chesser 1991b con FstatSim (Figura 11) son visualmente idénticas, las de Chesser 1991a con FStatSim (Figura 10) presentan leves diferencias. En este caso las de Chesser (Figura 10 (A)) tienen carencias de información, pues en los comentarios de las gráficas, y al momento de referirlas en el cuerpo de texto, omite los valores de d_f, b y ϕ usados para obtenerlas. Las diferencias que pueda haber en estas comparaciones se pueden atribuir a cuestiones de la escala usada, o limitaciones de resolución.

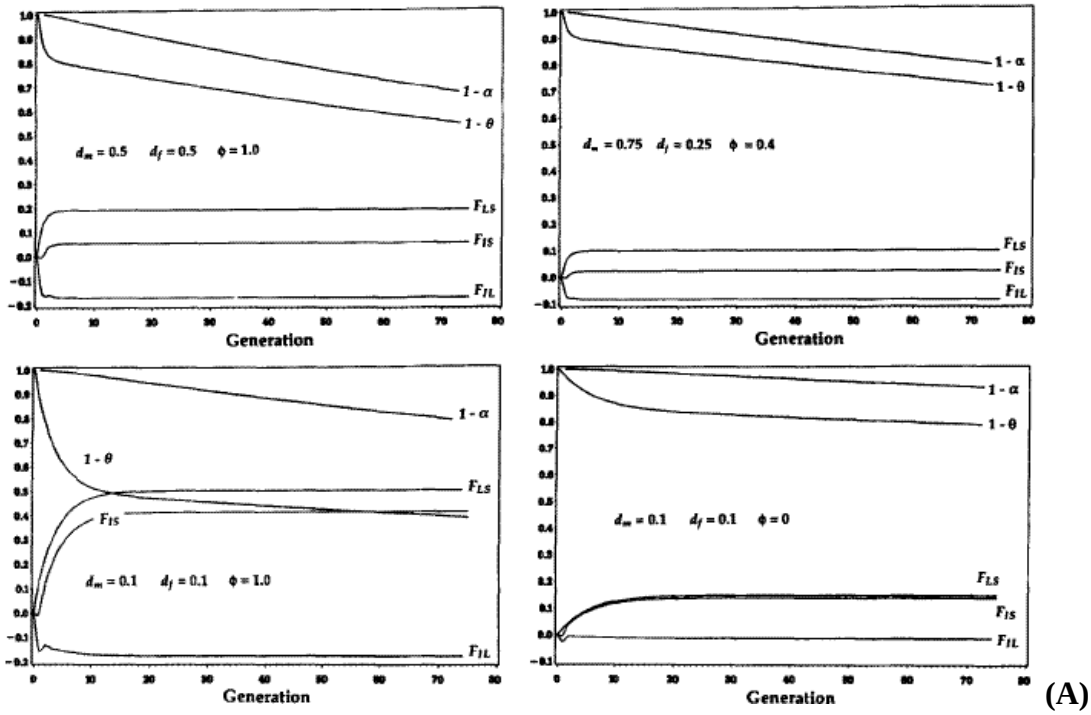


(A)

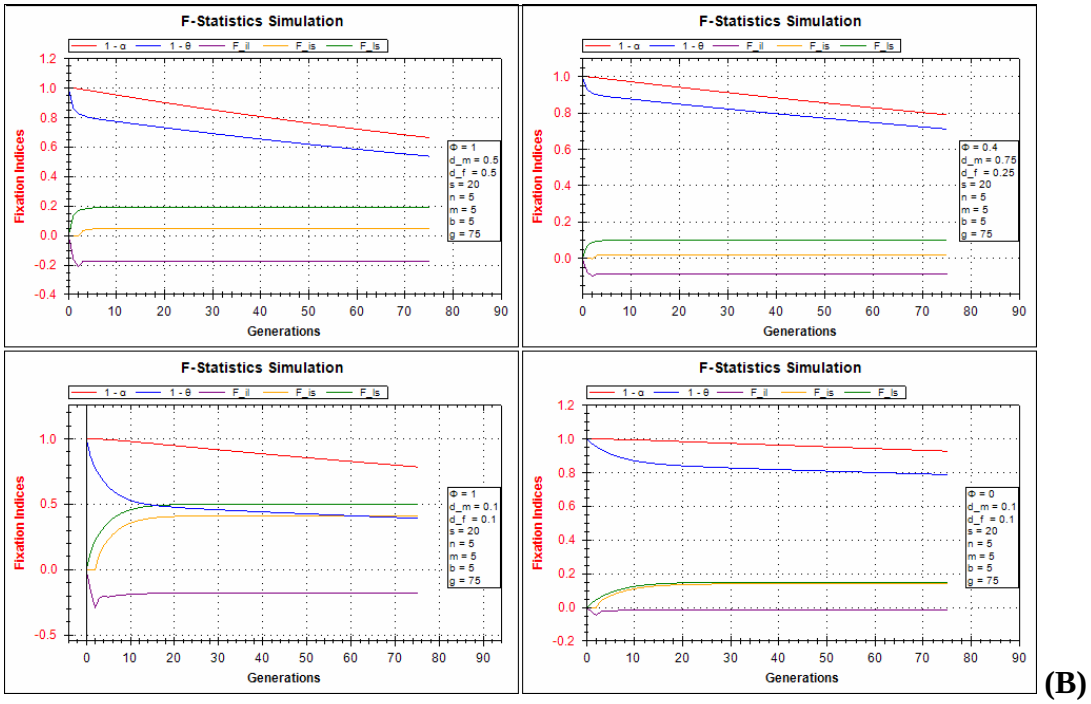


(B)

Figura 31 Comparación de las gráficas de Chesser 1991a(A) con las de FstatSim(B)



(A)



(B)

Figura 32 Comparación de las gráficas de Chesser 1991b(A) con las de FstatSim(B)

En varias pruebas con diferentes parámetros de entrada del programa, se observó que siempre que se replicaban en la matriz general T_3 las condiciones de la matriz particular T_1 (es decir, el caso d_m=1, d_f =0), las dos gráficas eran idénticas, mientras que al comparar de la misma manera T_3 con T_2 (el caso $d_m = d_f$ =1), ocurría lo mismo solo mientras se mantuviera ϕ=0. Cuando se colocaba el valor de ϕ en 1, ocurría una diferencia visible (aunque mínima: hasta de 0,03 unidades) entre las curvas graficadas con T_3 y T_2 para un mismo conjunto de datos iniciales. La tendencia se hacía más leve con valores de ϕ progresivamente más cercanos a 0, y la tendencia general era a disminuir el valor de las correlaciones genéticas y aumentar el de los índices de fijación al usar T_3, en relación con T_2.

Estos resultados, junto con el hecho de que Chesser 1991b presenta un análisis matemático más profundo que Chesser 1991a, se prestan para la conclusión de que en cualquier caso es preferible usar T_3 antes que T_2. Sin embargo, sería conveniente mantener la funcionalidad del programa para todos los casos, tanto para fines académicos como de referencia futura, hasta que el asunto se haya esclarecido mejor.

Como se ha mencionado, las discrepancias visuales entre los conjuntos de gráficos mostrados en la Figura 10 pueden deberse a cuestiones de la escala usada, o limitaciones de resolución por parte de Chesser. Un indicio que apunta a la primera posibilidad es el hecho de que en el caso de la gráfica inferior derecha de la Figura 10, al ser presentada para 500 generaciones (una diferencia exacta de un factor de 10) en FstatSim presenta una similitud considerablemente mayor con la gráfica correspondiente de Chesser. Otro indicio apunta a la segunda posibilidad, y es el hecho de que los picos al inicio de las gráficas de la Figura 11 se ven mucho mejor definidos (pero no más pronunciados) en las gráficas de FstatSim que en las de Chesser. En todo caso, las gráficas de Chesser fueron producidas hace más de 10 años en Fortran, (Chesser, comunicación personal) cuyo fuerte es la matemática, pero no la presentación visual, mientras que las de FstatSim fueron producidas en el presente, con una tecnología superior en gráficas, y una precisión matemática idéntica.

PRUEBAS DE CONCEPTO

La tabla 7 muestra el progreso, a través de tres versiones preliminares del programa, de tres criterios cualitativos de evaluación.

Tabla 7 Pruebas de concepto para diferentes versiones y criterios.

Versión \ Criterio	Apariencia	Facilidad de uso	Inteligibilidad
FStatSim V1 Beta1	50%	50%	25%
FStatSim V1 Beta2	100%	75%	50%
FStatSim V1 Final	100%	100%	100%

La apariencia es la agradabilidad visual del programa en aspectos de color y diseño visual. La facilidad de uso es qué tan fácil es de usar el programa, según la facilidad para ingresar los datos y obtener resultados a partir de los datos ingresados. La inteligibilidad evalúa si el software permite comprender cómo se ingresan los datos. Las encuestas se hicieron en cuatro usuarios familiarizados con el uso de software de manejo estadístico, a los cuáles se les preguntó si consideraban que el software era satisfactorio en cada uno de los criterios. La tabla muestra el porcentaje de usuarios que consideró satisfactorio el criterio dado para la versión dada.

COMPORTAMIENTO DE LAS GRÁFICAS

El comportamiento de las gráficas del modelo muestra que a pesar de una pérdida permanente de variación genética dentro de la población, los índices de fijación alcanzan rápidamente (en cerca de 10 generaciones) valores asintóticos. Además muestran que los valores de F_{IL} son negativos, indicando un exceso de heterocigosidad dentro de los linajes, y que los valores asintóticos de F_{LS} son inversamente proporcionales tanto al número de machos como al de hembras, por linaje y a la tasa de migración para cualquier sexo. Es lógico que con un número finito de linajes los valores de F y α se acumulen continuamente, erosionando la

variancia genética dentro de la población. Otro patrón es que el tiempo requerido para alcanzar la asíntota, aunque usualmente es veloz, es inversamente proporcional a la tasa de dispersión de ambos sexos. Para valores de ϕ cercanos a 1 junto con valores de dispersión de ambos sexos cercanos a 0, se experimentan los valores más altos de las asíntotas de F_{LS} y F_{IS}, mientras que la asíntota de F_{IL} es casi exclusivamente dependiente de ϕ (directamente proporcional). En general ϕ muestra tener una mayor influencia sobre los valores asintóticos de los tres estadísticos, que la tasa de dispersión de ambos sexos. Los valores de las asíntotas de los índices de fijación son medidas del cambio incremental del cambio en las correlaciones genéticas (por ejemplo, entre menos negativo es F_{IL}, menos negativa será la pendiente de las correlaciones genéticas).

La opción de FstatSim de cargar archivos de Genepop permite aplicar el modelo simulacional a datos biológicos reales, aunque estos deben usarse con precaución, pues el modelo asume ciertas condiciones que pueden no cumplirse en la población de donde se hayan extraído los datos. El modelo asume panmixia, es decir, que los individuos migrantes pueden moverse libremente entre las diferentes subpoblaciones de la población total, sin restricciones por distancias o barreras, con una disminución de diversidad genética y biodiversidad. No existe inmigración ni emigración, es decir, la población total se encuentra totalmente aislada de otras poblaciones. Se considera que la población ha sido recientemente originada, es decir, este aislamiento ocurrió en el momento en que fue muestreada la población. Todos los individuos se reproducen simultáneamente en cada generación, sin superposición de generaciones.

POSIBLES APLICACIONES

Una aplicación posible del programa es la de usar archivos de Genepop obtenidos de poblaciones biológicas silvestres con estructura social para graficar las cinco curvas y manipular los datos iniciales para tratar de obtener los valores de asíntota que más se asemejen a los valores iniciales (momentáneos, o de la generación 0) a partir de las

frecuencias genotípicas. Esto proveería una estimación informada de las tácticas reproductivas y tasas de dispersión de machos y/o hembras en la población muestreada, pues se parte del supuesto de que al ser una población silvestre, los estadísticos F estimados corresponden a los valores asintóticos alcanzados mediante una larga sucesión de generaciones anteriores manteniendo estas constantes poblacionales.

Otra aplicación posible consistiría en efectuar muestreos de campo sobre una población relativamente aislada y/o en peligro de extinción con estructura social, para tratar de predecir la rapidez con la que decaería la diversidad genética total al proponer programas de manutención de las poblaciones mediante transporte artificial de individuos de diferentes edades y sexos, desde y hacia grupos determinados, con miras a maximizar la retención de la diversidad genética, al evaluar qué combinaciones de valores tendrían mayor influencia en las pendientes de las curvas de correlaciones genéticas.

7. Conclusiones

El programa FstatSim produce gráficas a partir de datos ingresados por el usuario, ya sean digitados, o cargando un archivo de Genepop, para producir una gráfica que muestra las curvas de cinco valores que miden la diversidad genética de la población simulada. A nivel general, se observó en las gráficas que los estadísticos F alcanzaron asíntotas rápidamente, y que los valores de estas asíntotas no dependen tanto de los parámetros de coancestralidad iniciales, sino de los parámetros que describen las tácticas reproductivas y tasas de dispersión de la población. Estos valores son a su vez medidas de la acumulación incremental de correlaciones genéticas, por lo que se relacionan con la rapidez a la que estas curvas alcanzan su máximo valor asintótico, que es siempre uno, al tratarse de un modelo con poblaciones aisladas en las que un número finito de subpoblaciones panmícticas intercambian potencialmente indefinidamente su información genética.

Dada la ausencia de una aplicación que efectúe los cálculos que permite el modelo implementado en el programa desarrollado en el presente trabajo, y la necesidad de nuevos métodos que faciliten la comprensión visual y el procesamiento preciso de información biológica compleja, FstatSim puede ser una herramienta oportuna que podría llegar a tener cierta acogida en la comunidad científica. Las pruebas han indicado que la herramienta cumple los múltiples requisitos propios de un sistema informático validado, para iniciar su implantación y distribución entre los usuarios que puedan aprovechar sus atributos para expandir el cuerpo de conocimientos en el área de las ciencias biológicas.

8. Recomendaciones

Considerando que aproximadamente el 70% de la codificación del programa está dedicada a la lectura de archivos con el formato de Genepop, se recomienda continuar el trabajo mediante la implementación de nuevas funcionalidades que usen de más maneras la rica información biológica de los archivos Genepop.

Otra recomendación es la implementación de métodos matemáticos para obtener de manera rápida y confiable los conjuntos de datos necesarios para obtener una combinación de asíntotas deseadas con una pendiente deseada de la curva de correlaciones genéticas.

Una recomendación más es modificar el código para que sea compatible con el proyecto Mono, que permite la compilación de aplicaciones de Visual Basic .Net en plataforma Linux, lo cuál sería ideal para ampliar las plataformas disponibles, teniendo en cuenta que una porción considerable de la comunidad científica prefiere esta opción.

9. Referencias

Chesser, R. K. 1991a. Gene diversity and female philopatry. Genetics 127: 437-447.

Chesser, R. K. 1991b. Influence of gene flow and breeding tactics on gene diversity within populations. Genetics 129: 573-583.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution 23: 72-84.

Cockerham, C. C. 1973. Analysis of gene frequencies. Genetics 74: 679-700.

Contreras, R. 2006. Análisis y diseño de sistemas de información. El ciclo de vida del desarrollo de sistemas de información I. Instituto Tecnológico de Morelia. [en línea]: <http://deneb.itmorelia.edu.mx/cursos/mod/resource/view.php?id=18> [Fecha de consulta: 01/2007]. Pp 18-30.

Dix, A. J. 1991. Formal methods for interactive systems. Academic Press. London, England. 369 p.

Dobson, F. S., R. K.Chesser, J. L. Hoogland, D. W. Sugg, & D. Foltz. 1998. Breeding groups and gene dynamics in a socially structured population of prairie dogs. Journal of Mammalogy 79: 671-680.

Excoffier L. and G. Heckel. 2006. Computer programs for population genetics genetics data analysis: a survival guide. Nature Reviews Genetics. 7:745-758.

Gaskin, J. F., and B. A. Cabellero, and J. F. Crow. 2003. On the persistence and pervasiveness of a new mutation. Evolution 57: 2644-2646.

Hedrick, P. W. 2005. Genetics of Populations. Third Edition. Jones and Bartlett Publishers (eds.) Sudbury, Massachusetts, USA. 736 p.

Howard, R. D., J. A. DeWoody, and W. M. Muir. 2004. Transgenic male mating advantage provides opportunity for Trojan gene effect in a fish. Proc. Natl. Acad. Sci. USA 0101:2934-2938.

Lalouel, J., and N. E. Morton. 1973. Bioassay of kinship in a South American Indian population. Am J. Hum. Genet. 25: 62-73.

Law, A. M. & W. D. Kelton. 1991. Simulation modeling and analysis. Third Edition. McGraw Hill Higher Education (ed.) USA. 784 p.

Malecot, G. 1969. The Mathematics of Heredity. W. H. Freeman, San Francisco.

Morton, N. E., S. Yee, D. E. Harris and R. Lew. 1971. Bioassay of kinship. Theor. Pop. Biol. 2: 507-524.

Neel, J. V. and R. H. Ward. 1972. The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by F-statistics (including a comparison with the Makiritare and Xavante). Genetics 72L 639-666.

Nei, M. 1977. F-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 59: 327-332.

Nei, M. 1986. Definition and estimation of fixation indices. Evolution 40: 643-645.

Nei, M. and R. K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 41: 225-233.

Prout, T. 1981. A note on the island model with sex dependent migration. Theor. Appl. Genet. 59: 327-332

Raymond M. & Rousset F. 2005. Genepop on the web. README FILE. [en línea]: <http://genepop.curtin.edu.au/> [Fecha de consulta: 01/2007].

Rothman, E. D., C. F. Sing and A. R. Templeton. 1974. A model for analysis of population structure. Genetics 76: 943-960.

Vilá, C., A.-K. Sundqvist, O. Flagstand, J. Seddon, S. Bjornerfeldt, I. Kojola, A. Casulli, H. Sand, P. Wabakken, H. Ellegren. 2003. Rescue of a severely bottlenecked wolf (Canis lupus) population by a single immigrant. Proc. R. Soc. Lond. B. 270:91-97.

Weir, B. S. & C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Wright S. 1951. The genetical structure of populations. Annals of Eugenic 15: 323-54.

Wright, S. 1978. Evolution and the Genetics of Populations, Vol 4. Variability Within and Among Natural Populations. University of Chicago Press, Chicago.