Herramienta computacional para estimar la tasa de migración de flujos genéticos en diferentes poblaciones

C. G. Villegas[1], R. Muñíz-Salazar[2]*, D. L. Flores[3]

[1] Departamento de Sistemas y Computación, Instituto Tecnológico de Durango, Durango, México.
[2] Escuela de Ciencias de la Salud, Universidad Autónoma de Baja California, Ensenada, México.
[3] Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California, Ensenada, México.
*ramusal@uabc.edu.mx

Resumen

El flujo genético es la transferencia de alelos de genes de una población a otra, el cual influye en la variabilidad genética de una población y con ello de su evolución, dependiendo si hay mayor o menor flujo genético entre los individuos de las poblaciones. Por tal motivo, es importante, determinar tanto la dirección, como la intensidad del flujo genético entre las poblaciones. Actualmente, existen diversos programas de cómputo gratuitos que permiten calcular el flujo genético entre poblaciones. Uno de ellos y que es ampliamente usado es Migrate, con base en inferencia bayesiana. Sin embargo, este programa no realiza la interpretación de los resultados, por lo que, es necesario realizarlo de manera manual. Por ello, el presente trabajo trata sobre el desarrollo de una herramienta con interfaz gráfica basada en Python para estimar la tasa de migración de flujo genético entre poblaciones, aplicando la prueba t de Student no pareada con varianza desigual. La principal ventaja de esta herramienta es la automatización en la selección y extracción de datos, y en la aplicación de los cálculos estadísticos y con ello poder determinar si el flujo genético entre dos poblaciones es unidireccional o bidireccional.

Palabras clave: Flujo genético, Genética poblaciones, Migración, Migrate, Python.

1 Introducción

Una población es un grupo local de individuos que pertenece a una especie, además, es una entidad genética abierta, que puede intercambiar genes con otras poblaciones de la misma especie, mientras que la especie es una entidad cerrada, que no puede intercambiar genes con otras entidades. Las poblaciones son dinámicas; pueden crecer y expandirse o disminuir y contraerse mediante cambios en las tasas de nacimiento o mortalidad, o por migración o fusión con otras poblaciones. Las poblaciones raramente son sistemas cerrados. Por lo regular, se produce cierta cantidad de transferencia de genes, lo cual es más probable cuando las poblaciones se encuentran estrechamente relacionadas espacial y genéticamente. Entre las poblaciones adyacentes de una especie el flujo de genes puede ser grande, por lo que es de esperar que las poblaciones contiguas posean una composición génica más semejante que las que están más alejadas geográficamente. Por ello, la migración entre grupos geográficamente aislados es un suceso de gran importancia porque los complejos génicos se alteran y, en general, todas las diferencias genéticas entre las poblaciones se reducen [1].

La eficacia del intercambio de genes depende de la estructura de las dos poblaciones (emigrante y receptora) y, más específicamente, de la cantidad de migración (índice de migración, m) y de la magnitud de la diferencia en frecuencias génicas entre las dos poblaciones [1].

Actualmente, existen diferentes softwares de aplicación que describen la diversidad genética dentro y entre poblaciones [2], entre los más importantes se encuentran: Fstat, Genepop y Mígrate, centrándose en este último, Migrate [3] estima el tamaño de población efectivo, las tasas de migración pasadas

dx.doi.org/10.24254/CNIB.20.55

entre n poblaciones asumiendo un modelo de matriz de migración con tasas de migración asimétricas y diferentes tamaños de subpoblaciones, y divergencias o mezclas de poblaciones. Migrate utiliza la inferencia bayesiana para estimar conjuntamente todos los parámetros. Además, Migrate corre con datos de secuencias de DNA con o sin variación de velocidad del sitio, datos de polimorfismo de un solo nucleótido (SNP), datos de microsatélites y electroforéticos.

Sin embargo, el programa Migrate no determina si existen diferencias significativas entre poblaciones y la dirección del flujo, razón por la cual se propone el desarrollo de esta herramienta. Este trabajo presenta una herramienta gráfica para el procesamiento de datos resultantes de los cálculos realizados en el software Migrate, determina si existen diferencias significativas en el número de migrantes y la dirección del flujo genético entre poblaciones y el usuario puede seleccionar el nivel de confianza para realizar el cálculo. La ventaja principal de esta herramienta gráfica es el ambiente amigable e intuitivo para el usuario.

2 Metodología

2.1 Conjunto de datos

Se parte de un archivo de salida (outfile) que es generado por Migrate, el cual contiene una tabla con el número de migrantes (MLE) por población calculados con diferentes percentiles. En la Fig. 1 se muestra un ejemplo de la tabla generada por Migrate, en el que se muestran seis columnas que se describen a continuación.

- Columna 1. Parámetro. Muestra los valores de $Theta$ para cada n población, y posteriormente muestra el par de población analizada. M_21, corresponde al par de la población 2 y 1, y así sucesivamente.
- Columnas 2 a 5. Percentiles bajos para cada parámetro.
- Columna 6. MLE. Corresponde al valor de número de migrantes para cada par de población.

2.2 Extracción de datos

Para extraer solo los datos necesarios del archivo de salida de Migrate, se identifican los renglones clave dentro de los archivos, los cuales a partir de su ubicación contendrán la información necesaria. Es importante analizar el formato del contenido, ya que es posible encontrar patrones que faciliten la ubicación de los parámetros deseados. Dentro de los archivos de salida de Migrate se encuentran cinco columnas clave:

1. El inicio del total de poblaciones con su respectivo total individual
2. El fin del total de poblaciones
3. Los percentiles bajos: 0.005, 0.025, 0.050 y 0.250
4. El valor de MLE
5. Los percentiles altos: 0.750, 0.950, 0.975 y 0.995

Como se observa en la Fig. 1, la tabla contiene valores de $Theta$, $(Theta\_1, Theta\_2, \dots)$, dichos valores no son necesarios para el análisis, por lo que se deben omitir. Sin embargo, con esta información, se puede conocer la cantidad de pares a analizar, en este ejemplo, se tienen 14 pares ($Theta\_14$).

Para conocer la cantidad de pares a analizar (parámetros), se hace uso del cálculo de permutaciones, como se muestra en la ecuación (1).

$$P_r^n = \frac{n!}{(n-r)!} \tag{1}$$

donde: n es el total de poblaciones y r es el par de poblaciones. Para este caso, dado que se esta trabajando con pares, entonces $r = 2$.

```
=============================================================================
Summary of profile likelihood percentiles of all parameters
=============================================================================
```

Parameter	Lower percentiles				
	0.005	0.025	0.050	0.250	MLE
Theta_1	0.424302	0.436666	0.443169	0.464086	0.479349
Theta_2	0.299830	0.311888	0.318304	0.339206	0.354828
Theta_3	0.468880	0.483731	0.491571	0.516801	0.535359
Theta_4	0.361483	0.374756	0.381804	0.404668	0.421622
Theta_5	0.472815	0.491434	0.501313	0.533496	0.557550
Theta_6	0.617537	0.637129	0.647487	0.680819	0.705338
Theta_7	0.380955	0.396905	0.405412	0.433235	0.454031
Theta_8	0.489382	0.505043	0.513297	0.539956	0.559567
Theta_9	0.830797	0.888313	0.919882	1.026782	1.111256
Theta_10	0.578636	0.598668	0.609279	0.643571	0.668999
Theta_11	0.328612	0.352193	0.365124	0.409267	0.444149
Theta_12	0.275332	0.290882	0.299305	0.327335	0.348963
Theta_13	0.647592	0.673326	0.687019	0.731629	0.764971
Theta_14	0.498433	0.519675	0.531021	0.568002	0.595865
M_21	144.188036	161.211053	170.453597	201.181520	224.517847
M_31	4.97477e-11*	7.46165e-11*	8.7051e-11*	9.32682e-11*	9.94854e-11
M_41	182.605054	223.807209	247.275161	329.915700	397.317557
M_51	1.0077e-10*	1.5115e-10*	1.7634e-10*	1.88935e-10*	2.0153e-10
M_61	1.34102e-10*	2.01148e-10*	2.34671e-10*	2.51433e-10*	2.68194e-10
M_71	61.100873	90.871672	109.604802	184.611118	253.437138
M_81	1.89217e-10*	2.8382e-10*	3.31122e-10*	3.54773e-10*	3.78424e-10
M_91	265.138865	340.702634	384.775498	545.089765	680.030551
M_101	2.43747e-10*	3.65615e-10*	4.26549e-10*	4.57016e-10*	4.87483e-10
M_111	2.82487e-10*	4.23726e-10*	4.94345e-10*	5.29654e-10*	5.64964e-10

Fig. 1. Ejemplo de un archivo generado por Migrate

2.3 Procesamiento de datos

Una vez que se extraen los datos del archivo de salida de Migrate, estos son almacenados en estructuras tipo listas, vectores y matrices de Python [4]. En la Fig. 2 se muestra una sección del código donde observa la forma en la que se obtuvieron los datos para su procesamiento en la interfaz.

El desarrollo de este trabajo fue realizado en Python v.3.8.4 haciendo uso de sus librerías [5].

1. Para creación de interfaz gráfica, *PyQt5*
2. Para aplicar prueba t de Student, *scipy*
3. Para selección, extracción y manipulación de datos, *Pandas*

Es justo aquí donde inicia la automatización del cálculo estadístico entre pares de parámetros o poblaciones. Por ejemplo, el parámetro **M_21** significa que el flujo genético va de la población 2 a la población 1 y por su parte el parámetro **M_12** significa que el flujo genético va de la población 1 a la población 2.

Los pares de poblaciones se deben comparar estadísticamente por medio de una prueba de t de Student, ecuación (2), para determinar si existen diferencias significativas en la dirección del flujo genético entre poblaciones. Si hay diferencia significativa, entonces una población es la que está enviando mayor número de individuos (migrantes) a la otra población. Si no hay diferencia significativa, entonces el flujo es bidireccional.

$$t_{\sigma/2} = \frac{\mu_1 - \mu_2}{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{2}$$

```
mixed_percentiles = self.percentiles_df[['Parameter','0.005','0.995','MLE']]
#mixed_percentiles['Sum_of_values'] = mixed_percentiles.sum(axis=1)

first_element = list(combinations((self.total_per_group), len(self.total_per_group)-1))
first_element.reverse()

first_element_list = []

for tuple_element in first_element:
    for list_element in tuple_element:
        first_element_list.append(list_element)

second_element_list = [element for element in self.total_per_group for i in range(len(self.total_per_group)-1)]

mixed_percentiles['First_total'] = first_element_list
mixed_percentiles['Second_total'] = second_element_list

print(mixed_percentiles)

n_permutations = len(list(permutations(range(len(self.total_per_group)),2)))
steps = len(self.total_per_group) - 1
start = len(self.total_per_group) - 1
second_pair = []

for i in range(0, steps):
    start = i * steps + steps + i
    for j in range(start, n_permutations, steps):
        second_pair.append(j)
```

Fig. 2. Código en Python para el procesamiento de los datos obtenidos de Migrate

donde μ_1 y μ_2 son las medias de los dos conjuntos de datos muestra, σ_1 y σ_2 son las varianzas de las dos poblaciones y n_1 y n_2 son los números de elementos de cada muestra.

Para llevar a cabo el análisis estadístico, se contemplan las hipótesis, pruebas de decisión e interpretaciones que se muestran en la Tabla 1. La Fig. 3 muestra gráficamente estas pruebas de hipótesis y su interpretación. MLE21 representa el número de migrantes de la población 2 a la población 1, MLE12 indica el número de migrantes de la población 1 a la población 2, t_{tab} es el valor del estadístico t obtenida de tablas estadísticas y t_{calc} es valor del estadístico t obtenido de la ecuación (2).

Tabla 1: Prueba de hipótesis y su interpretación

Hipótesis	Decisión	Interpretación
H_o : MLE21 = MLE12	$-t_{tab} > t_{calc} > t_{tab}$ se rechaza H_o	Flujo genético es unidireccional
H_A : MLE21 $\neq$ MLE12	$-t_{tab} < t_{calc} < t_{tab}$ se acepta H_o	Flujo genético es bidireccional

3 Resultados y Discusión

La interfaz principal es intuitiva, como se muestra en la Fig. 4, se encuentran cuatro botones que guían al usuario a realizar el análisis estadístico entre pares de poblaciones de manera sencilla.

Al oprimir el botón *Open file* se abre un directorio de archivos que permite al usuario elegir un documento con extensión *.txt*, al seleccionar el archivo en la cabecera de la interfaz se muestra por medio de etiquetas la dirección en la que se encuentra el archivo y el número de poblaciones analizadas en Migrate. Durante el proceso de carga de los datos, se crean tres archivos con extensión *.csv*, el primero contiene las poblaciones con su respectivo total individual y los restantes contienen los percentiles. El archivo de poblaciones obtiene el total de las mismas y los archivos con percentiles se unen para almacenar y mostrar sus datos en una tabla. El botón *Run* muestra la tabla generada en el proceso

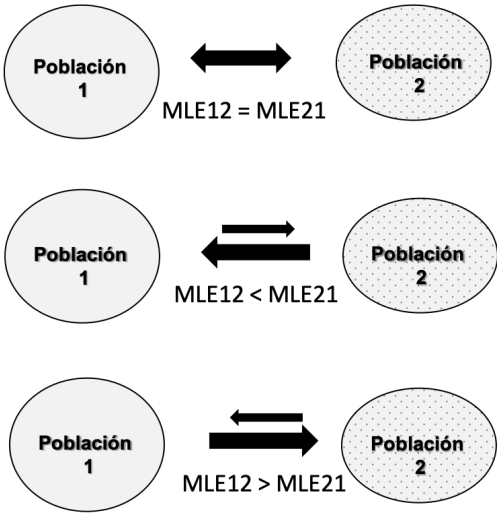

Fig. 3. Número de migrantes que dona la población 1 a la población 2 que es receptora (MLE12) y viceversa (MLE21). Las flechas indican la dirección y la intensidad del flujo genético. Las flechas de mayor tamaño y con una sola dirección indican que una de las dos poblaciones es la que está donando una mayor cantidad de migrantes a la población receptora.

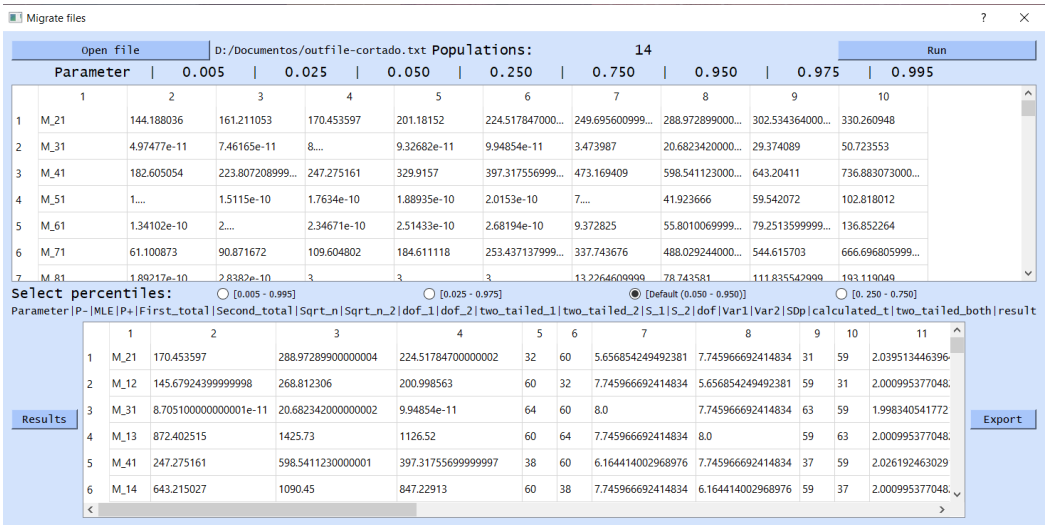

Fig. 4. Interfaz principal

anterior.

Para la selección de valores con los que se quiera realizar el análisis, en la parte intermedia de la interfaz se presentan diferentes opciones que representan los percentiles, el usuario solo puede escoger una de ellas, al escoger, el botón *Results* realiza los cálculos con los valores de esas columnas, aplica la prueba t de Student, y muestra los resultados en otra tabla.

El botón *Export* permite al usuario exportar los resultados anteriores a un archivo con extensión *.csv*, para ser visualizados como tabla, como se observa en la Fig. 5.

	A	B	C	D	E		R	S	T	U
1	Parameter	0.05	0.95	MLE	First_total		SDp	calculated_t	two_tailed_t	result
2	M_21	170.453597	288.972899	224.517847	32		51820.0409	0.46575853	1.98667454	Unidirectional
3	M_12	145.679244	268.812306	200.998563	60		57361.336	3.78328381	1.98667454	Unidirectional
4	M_31	8.71E-11	20.682342	9.95E-11	64		6640.06707	-76.307819	1.97959988	Unidirectional
5	M_13	872.402515	1425.73	1126.52	60		1389713.4	3.41430142	1.97959988	Unidirectional
6	M_41	247.275161	598.541123	397.317557	38		517352.044	-2.9828082	1.98498431	Unidirectional
7	M_14	643.215027	1090.45	847.22913	60		755840.506	4.64704684	1.98498431	Unidirectional
8	M_51	1.76E-10	41.923666	2.02E-10	42		22959.9583	-13.779539	1.98397152	Unidirectional
9	M_15	240.185875	685.270857	424.524068	60		888158.605	2.21551649	1.98397152	Unidirectional
10	M_61	2.35E-10	55.801007	2.68E-10	56	...	45095.1835	5.18E-12	1.9809923	Unidirectional
11	M_16	5.42E-11	17.960341	6.20E-11	60		4671.70318	-19.782791	1.9809923	Unidirectional
12	M_71	109.604802	488.029244	253.437138	36		696609.794	-0.4225617	1.98552344	Unidirectional
13	M_17	106.483341	746.468358	328.684074	60		2209362.54	1.03643309	1.98552344	Unidirectional
14	M_81	3.31E-10	78.743581	3.78E-10	54		88331.6257	-5.15E-12	1.98137181	Unidirectional
15	M_18	5.85E-10	21.666636	6.68E-10	60		6687.57823	-43.938873	1.98137181	Unidirectional
16	M_91	384.775498	1097.71	680.030551	16		2208867.65	-0.6463503	1.99254349	Unidirectional
17	M_19	606.385895	1424.18	957.808022	60		2753901.33	1.9959978	1.99254349	Unidirectional
18	M_101	4.27E-10	101.394512	4.87E-10	56		148893.241	-2.11E-09	1.9809923	Unidirectional
19	M_110	1.34E-07	82.097474	1.53E-07	60		97612.5535	2.61E-09	1.9809923	Unidirectional
20	M_111	4.94E-10	117.532085	5.65E-10	16		174902.639	2.11E-12	1.99254349	Unidirectional
21	M_111	2.71E-10	191.512165	3.10E-10	60		464382.893	-1.47E-12	1.99254349	Unidirectional

Fig. 5. Archivo que se genera con la herramienta computacional desarrollada en donde las poblaciones se encuentran ordenadas por pares de poblaciones de manera consecutiva y muestra el tipo de migración entre poblaciones.

4 Conclusiones

En el presente trabajo se mostró una herramienta con interfaz gráfica para estimar la tasa de migración de flujos genéticos en diferentes poblaciones.

Está desarrollada en Python y la interfaz es amigable con el usuario ya que el flujo de los elementos en la interfaz es fácil de seguir, y es posible usarla sin necesidad de saber programar, sin descargar un entorno de programación y sin conocer el lenguaje. Como trabajo futuro, se pretende subir la aplicación a un servidor web, de esta manera será más accesible su uso para estudiantes, maestros e investigadores.

Declaración de conflictos de interés

Los autores declaran no tener ningún conflicto de interés para este trabajo.

Agradecimientos

Los autores desean agradecer a la Academia Mexicana de Ciencias por haber otorgado una beca a la estudiante Cinthya-Guadalupe Villegas para su estancia de investigación de verano virtual.

Referencias

[1] A. Templeton, *Gene Flow and Subdivided Populations*, ch. 6, pp. 155–193. Academic Press, 2019.

[2] T. Garrido-Garduño and E. Vázquez-Domínguez, "Métodos de análisis genéticos, espaciales y de conectividad en genética del paisaje," *Biodivers*, vol. 84, no. 3, pp. 1031–1054, 2013.

[3] P. Beerli, *MIGRATE: documentation and program. (Version 4.0)*. Washington: Department of Scientific Computing, Florida State University, 2016.

[4] M. Summerfield, *Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming*. Prentice Hall, 1st ed., 2006.

[5] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, and C. SciPy, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, 2020.