UNIVERSIDAD DE CHILE

Tasa de evolución molecular y pleiotropía en la trayectoria evolutiva de secuencias codificantes

Tesis entregada a la Universidad de Chile en cumplimiento parcial de los requisitos para optar al grado de:

Doctor en Ciencias con mención en Ecología y Biología Evolutiva

Facultad de Ciencias

Por

Felipe Ignacio Avello Duarte

Director de tesis: Dr. Elie Poulin
Co-director de tesis: Dr. Pablo Razeto Barry

Abril, 2024

FACULTAD DE CIENCIAS

UNIVERSIDAD DE CHILE

INFORME DE APROBACIÓN

TESIS DE DOCTORADO

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Doctorado presentada por el candidato

Felipe Ignacio Avello Duarte

Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Doctor en Ciencias con mención en Ecología y Biología Evolutiva, en el examen de Defensa Privada de Tesis rendido el día _____

Director de tesis:

Dr. Elie Poulin _____

Co-director de tesis:

Dr. Pablo Razeto Barry _____

Comisión de evaluación de la tesis:

Dr. David Véliz _____

Dr. Marco Méndez _____

Dra. Karina Vilches _____

Dr. Rodrigo Vásquez _____

A mis profesores,
Lorens y Patricio

Resumen Biográfico

De chico, como a muchos otros niños y niñas, me gustaban los dinosaurios, la pasaba memorizando sus nombres y modos de vida, mirando por horas imágenes en revistas y libros sobre esos animales del pasado. Luego me hice fanático de todos los juegos y series que se vincularan a la diversidad y evolución de los seres vivos y pasé todas las clases de básica dibujando animales, criaturas fantásticas y sus transformaciones. Recuerdo con especial cariño *El Futuro es Salvaje*, un documental que me reveló a los 10 años la belleza de la evolución y el poder de creación de nuevas formas que tienen el tiempo y los organismos.

A diferencia de otros niños y niñas, jamás superé esa etapa, así que el último año de media decidí seguir mi pasión y entré a estudiar biología en la Universidad de Chile, y puedo decir con tranquilidad y alegría que tomé la decisión correcta. Con contadas excepciones disfruté cada clase, terminando todos los días asombrado con la complejidad y la belleza de los seres vivos, de su funcionamiento, sus interacciones y su historia. A lo largo del pregrado pasé por laboratorios de inmunología, biología del desarrollo y neurobiología, intentando encontrar mi lugar. Fue en este último donde realmente entendí lo que significa ser biólogo y vivir una vida de investigación. Luego de cursar Evolución el último semestre tomé la decisión de seguir esa línea que, además de encontrarla fascinante, me permitía compatibilizar la biología con mi interés por las matemáticas y la computación. Gracias al apoyo de mi tutor y co-tutor de tesis pude entrar y cursar este postgrado en biología evolutiva y desarrollar el presente trabajo de investigación, haciendo un pequeño aporte a nuestro entendimiento sobre los procesos de transformación histórica de los organismos. Esta tesis me acompañó y creció conmigo en un intenso periodo de turbulencia política, pandémica y emocional y mirando para atrás sólo puedo estar agradecido y orgulloso del camino recorrido.

Agradecimientos

El presente trabajo de investigación es el resultado de un proceso colectivo, donde cada persona que participó contribuyó con acciones invaluables tanto desde el conocimiento científico como desde el amor y el apoyo. En consecuencia, quisiera hacer un reconocimiento no exhaustivo de sus aportes, comenzando por mi director de tesis Elie Poulin, que me recibió con los brazos abiertos y me dio su apoyo y un espacio para desarrollar mis intereses. Agradecer a mi co-director de tesis Pablo Razeto, sin el cual no estaría acá, gracias por las conversaciones evolutivas tan enriquecedoras. Además, quisiera agradecer también a Alexia Núñez, que me acompañó en los primeros pasos que di en la ciencia y que ha estado ahí para apoyarme y seguir creciendo en este trabajo. Finalmente, agradecer a las y los tremendos investigadores que nutrieron mi tesis con sus conocimientos y apoyo, Felipe Urcelay, Claudio Hernández, Amitai Linker, Diego Pacheco, Juliana Vianna, Daly Noll y Luis Pertierra.

Quisiera agradecer a todas y todos los grandes formadores que he tenido dentro del aula, fueron un ejemplo a seguir y despertaron en mí el amor por el conocimiento y la docencia. No puedo ser lo suficientemente enfático en lo mucho que le agradezco a mis amistades, partiendo por las personas maravillosas que conocí en la facultad y con las cuales aprendí y gocé el estudio de las ciencias biológicas. Agradezco a mis compañeras y compañeros de organización, han sido una inspiración para seguir trabajando por la transformación radical de nuestro país y del mundo. Agradezco también a cada uno de los miembros del Laboratorio de Ecología Molecular, gracias por acompañarme con apoyo, risas y cervezas en los últimos pasos de la tesis, fueron el catalizador que necesitaba. También a las personas maravillosas que han llegado a mi vida estos últimos años y a aquellas que recuperé luego de tanto tiempo distanciados, les estoy infinitamente agradecido. Por último, a mis amigas y amigos del colegio que tantas veces fueron mi motivo, mi pañuelo de lágrimas, mi fuente de risas y de enojos y con los que espero seguir compartiendo el vivir, gracias por tanto.

Quiero expresar mi gratitud a mi familia y a quienes me han querido como a uno más de sus familias, no pude haber escogido un lugar con más amor que aquel en el que crecí. Gracias a Cecilia Torres que me entregó el don de la constancia y la disciplina, no estaría donde estoy si no fuera por sus enseñanzas. Finalmente, gracias a Juanchi, Simón, Osita y a mi mamá Lorens y mi papá Patricio, que han aguantado mis llantos y mi mal genio, me han enseñado a querer de forma sincera, y a los cuales les debo todo lo que soy, son de lo más preciado que tengo, muchas gracias por todo.

Para terminar, agradezco a las instituciones que financiaron este proyecto, a la Facultad de Ciencias de la Universidad de Chile, a ANID y al Instituto Milenio BASE.

Índice de Materias

Lista de Figuras

Lista de Símbolos, Abreviaturas o Nomenclatura

- ADN: ácido desoxirribonucleico
- tARN: ácido ribonucleico de transporte
- mARN: ácido ribonucleico mensajero
- DFE: distribution of fitness effects
- FGM: Fisher's geometric model
- PPI: protein-protein interaction
- MSD: mean squared displacement
- Pe: Péclet number
- KS: Kolmogorov-Smirnov
- GO: Gene Ontology
- CDS: coding sequence
- NCBI: National Center for Biotechnology Information
- nBLAST: nucleotide Basic Local Alignment Search Tool
- HOG: hierarchical orthologous group
- PAML: Phylogenetic Analysis by Maximum Likelihood
- d_N: number of non-synonymous substitutions divided by the number of non-synonymous sites
- d_S: number of synonymous substitutions divided by the number of synonymous sites
- Mya: million years ago
- GBIF: Global Biodiversity Information Facility

- GOEA: Gene Ontology enrichment analysis
- PCA: principal component analysis
- HPD: highest posterior density
- β: slope of the linear regression between evolutionary rate and pleiotropy

1 Resumen

Las investigaciones sobre las variables que modifican la tasa de cambio evolutivo de las macromoléculas han estado en el centro del estudio de la evolución molecular desde sus inicios. En esta línea la pleiotropía, entendida como la capacidad de una mutación de afectar múltiples rasgos fenotípicos, ha sido tratada como un obstáculo para la acumulación de nuevas mutaciones, disminuyendo la tasa evolutiva, debido a que aumenta la probabilidad de que la mutación sea deletérea, a la vez que disminuye su probabilidad de fijación. Si bien esto ha sido demostrado a un nivel teórico bajo condiciones específicas, aún no hay una respuesta satisfactoria de por qué no se ha encontrado esta clara relación negativa en la naturaleza ni tampoco se sabe cómo diferentes condiciones y supuestos pueden cambiar este patrón. Aquí hacemos uso del modelo geométrico de Fisher para estudiar en profundidad los efectos de la pleiotropía sobre la tasa evolutiva e incorporar, por una parte, un seguimiento temporal del proceso evolutivo y, por otra, el rol de la selección natural y la deriva genética sobre estos efectos, haciendo uso de una descripción del modelo basado en ecuaciones diferenciales parciales. Adicionalmente, evaluamos las predicciones del modelo a la luz la evolución de las secuencias codificantes de 15 especies de pingüinos y expandimos nuestros resultados para encontrar procesos biológicos significativos en la evolución de estas aves. En el modelamiento, así como en el análisis genómico, encontramos que el efecto de la pleiotropía sobre la tasa evolutiva cambia dinámicamente a lo largo de una trayectoria evolutiva siguiendo tres etapas estereotípicas un efecto negativo inicial, una segunda etapa de efecto positivo de mayor duración y una etapa final de efecto levemente negativo.

2 Abstract

Research on the variables that modify the rate of evolutionary change of macromolecules has been at the center of the study of molecular evolution since its inception. In this line pleiotropy, understood as the capacity of a mutation to affect multiple phenotypic traits, has been treated as a hindrance to the accumulation of new mutations, decreasing the evolutionary rate, because it increases the probability of the mutation being deleterious while decreasing its probability of fixation. While this has been demonstrated at a theoretical level under specific conditions, there is still no satisfactory answer as to why this clear negative relationship has not been found in nature, nor is it known how different conditions and assumptions can change this pattern. Here we make use of Fisher's geometric model to study in depth the effects of pleiotropy on the evolutionary rate and incorporate, on the one hand, a temporal follow-up of the evolutionary process and, on the other hand, the role of natural selection and genetic drift on these effects, making use of a description of the FGM based on partial differential equations. In addition, we evaluated the model predictions in light of the evolution of the coding sequences of 15 penguin species and expanded our results to find significant biological processes in the evolution of these birds. In our model, as well as in genomic analysis, we found that the effect of pleiotropy on evolutionary rate changes dynamically along an evolutionary trajectory following three stereotypic stages an initial negative effect, a second stage of longer-lasting positive effect, and a final stage of slightly negative effect.

3 Introducción general

3.1 Fundamentación teórica

> *"La historia evolutiva de los seres vivos es una historia de conservación y cambio, tanto de los linajes de organismos, como de los linajes de sistemas que se intersectan con ellos en su realización estructural"*(Maturana and Mpodozis, 1992)

La participación de las proteínas en los procesos orgánicos es ubicua, por lo que el cambio histórico de estas moléculas y de su sistema de herencia, mediado en parte por la estructura de las macromoléculas de ADN, son de gran interés en el estudio de la evolución biológica. Durante el devenir evolutivo de las especies ocurren múltiples eventos mutacionales de carácter contingencial en sitios del ADN portado por individuos dentro de una especie. Si las mutaciones ocurren en moléculas conservadas durante la reproducción, ya sea de forma directa o por una relación de continuidad mediada por la replicación del ADN, la reproducción de los organismos va a facilitar el proceso de propagación de la forma mutante dentro de la especie. La propagación va a verse afectada por múltiples factores que se pueden dividir en dos grandes grupos, aquellos factores que son independientes de la mutación y están vinculados a fluctuaciones aleatorias durante la ontogenia de los organismos (Wright, 1955) y aquellos factores que dependen del efecto sobre el fenotipo asociado a la mutación y están vinculados a modificaciones en

la sobrevida y reproducción (Darwin and Wallace, 1858), estos grupos están asociados a los conceptos de deriva genética y selección natural, respectivamente. Existen tres resultados posibles del proceso de propagación, en primer lugar, puede ocurrir, con alta probabilidad que la mutación se pierda, en segundo lugar, puede ocurrir que la mutación permanezca en una coexistencia estable con la forma ancestral, es decir, la mantención de un polimorfismo dentro de la especie, o, como tercera opción, la forma mutante puede reemplazar completamente a la forma ancestral en la población, evento llamado fijación o sustitución (Kimura, 1962). Las diferencias que observamos entre las secuencias de especies actuales son el resultado de múltiples eventos de fijación mediados por la selección natural y la deriva genética.

Se habla de una mutación beneficiosa, deletérea o neutral en relación al efecto que esta tiene sobre las tasas de natalidad y mortalidad de los organismos portadores al compararlas con las de aquellos que poseen la forma nucleotídica original (Doebeli et al., 2017), lo cual se asocia tradicionalmente al concepto de fitness. Naturalmente, que sea beneficiosa, deletérea o neutral no es una propiedad exclusivamente propia de la mutación, sino que es también determinantemente dependiente del contexto en el que ocurre. Excepciones a esto último se pueden encontrar, por ejemplo, en las mutaciones que ocurren en sitios sinónimos de las secuencias codificantes (CDS). Estas mutaciones tienen una mayor probabilidad de tener un efecto neutral independiente del contexto en el que ocurran debido a la degeneración del código genético donde modificaciones en la estructura nucleotídica no van a generar modificaciones en la estructura primaria de las proteínas, a pesar de que sí pueden tener efecto sobre otros elementos como el nivel de expresión o su estructura terciaria debido a la concentración particular de cada tipo de tARN, por ejemplo (Bailey et al., 2021). Que una mutación sea beneficiosa, deletérea o neutral va a afectar su probabilidad de fijación por medio, únicamente, de procesos selectivos, ya que los procesos de deriva genética son insensibles a tales propiedades. A

nivel teórico y empírico se ha estudiado en detalle la distribución de estas propiedades sobre todas las nuevas mutaciones que pueden ocurrir en un genoma, conocida como la distribución de efectos sobre el fitness (DFE), intentando identificar la proporción de sustituciones que son beneficiosas, deletéreas y neutrales (García-Dorado et al., 1998; Vale et al., 2012; Rice et al., 2015; Joyce and Abdo, 2018; Charmouh et al., 2023, para ver algunos ejemplos).

Una de las razones por las cuales conocer la DFE particular de una CDS es tan importante es debido a su efecto sobre la tasa evolutiva molecular (Wang and Zhang, 2009), el ritmo al cual las secuencias acumulan mutaciones. Se ha identificado un gran número de factores capaces de modificar la DFE de las CDS como, por ejemplo, la localización subcelular, la posición en la red molecular o la estabilidad del producto proteico de la CDS. Algunos de estos elementos característicos pueden enmascarar el efecto fenotípico de las mutaciones, aumentando la proporción de mutaciones neutrales, mientras que lo contrario ocurre con elementos que potencien el efecto de las mutaciones, donde van a aumentar las mutaciones beneficiosas y deletéreas a expensas de las neutrales. Si bien la DFE depende de varias variables como la complejidad del organismo y el contexto genético, por nombrar algunos (Bataillon and Bailey, 2014), la proporción de mutaciones deletéreas suele ser mucho mayor que la proporción de mutaciones beneficiosas, por lo que los casos de selección natural purificadora (que disminuye la tasa evolutiva) van a ser más comunes que los casos de selección natural direccional (que aumenta la tasa evolutiva). Consecuentemente, aquellos factores que enmascaran los efectos fenotípicos de las mutaciones, como la participación de chaperonas, un alto contenido de puentes de disulfuro o la presencia de una CDS idéntica, por ejemplo, van a tener el efecto de acelerar la tasa evolutiva, ya que van a aumentar la cantidad de mutaciones neutrales, aumentando la probabilidad de fijación promedio, lo que se puede pensar como una disminución en el poder de la selección natural purificadora. Por ejemplo, los eventos de

duplicación génica, ya sea tanto bajo escenarios de neofuncionalización como de subfuncionalización, evidencian una relajación de la selección natural purificadora gracias a que los cambios en el funcionamiento de una secuencia producto de una mutación pueden ser cubiertos por la copia (Pegueroles et al., 2013). Lo contrario ocurre con elementos que potencien el efecto fenotípico de las mutaciones como, por ejemplo, el nivel de expresión génica (Pál et al., 2001). Por múltiples motivos, pueden ocurrir errores de plegamiento durante la traducción del mARN, donde una proteína mal plegada puede generar interacciones proteína-proteína (PPI) no presentes en la proteína silvestre o, en mayores concentraciones, puede generar citotoxicidad. Existen mutaciones que aumentan la probabilidad de que ocurran errores en el plegamiento, causando un efecto deletéreo para el organismo. En CDS que tienen un alto nivel de expresión, esto implicaría un incremento de la concentración de proteínas mal plegadas, causando mayores niveles de citotoxicidad. Esta es una de las hipótesis por las cuales se piensa que existe una relación negativa tan marcada entre la tasa evolutiva y el nivel de expresión (Zhang and Yang, 2015).

Si bien la primera reconciliación entre la genética Mendeliana y la biología evolutiva Darwiniana fue realizada a comienzos del siglo XX por los primeros mutacionistas (Stoltzfus and Cable, 2014), fueron los fundadores de la genética de poblaciones los que asentaron las bases del estudio evolutivo como lo conocemos actualmente. Su aproximación fue teórica y matemática en naturaleza, en un contexto donde la gran cantidad de datos necesarios para generar un correlato empírico era inconcebible (Casillas and Barbadilla, 2017). En ese contexto es en el que Fisher, en un par de párrafos, describe un modelo (FGM) que permitió desarrollar una intuición sobre la relación entre la magnitud del efecto fenotípico de una mutación y la probabilidad de que sea beneficiosa (Fisher, 1930). Casi un siglo después ese modelo se ha desarrollado extensamente, con aplicaciones a una multitud de problemáticas evolutivas y con múltiples evaluaciones empíricas de sus

predicciones (Rice, 1990; Hartl and Taubes, 1996; Orr, 2000; Wilke and Adami, 2001; Martin et al., 2007; Gros et al., 2009; Manna et al., 2011; Lourenço et al., 2011; Blanquart et al., 2014; Hwang et al., 2017; Moutinho et al., 2022, para ver algunos ejemplos). La definición de la DFE es fundamental para muchos de los modelos evolutivos actuales y el FGM no es la excepción, sin embargo, a diferencia de otros modelos (véase por ejemplo Kingman, 1978), en el FGM la DFE no requiere ser especificada con anterioridad, sino que emerge de los parámetros del modelo. Consecuentemente, este modelo ha permitido ver el efecto de ciertas variables de interés evolutivo sobre la DFE. Sin ir más lejos, la primera utilización del modelo, realizada por su autor en 1930, fue la de identificar el cambio en la probabilidad de que una mutación sea beneficiosa en función de su magnitud de efecto fenotípico, encontrando que la probabilidad es 0.5 cuando el efecto tiende a cero y decayendo rápidamente a medida que la magnitud del efecto de la mutación aumenta. Una conclusión relacionada y que ha sido extensamente confirmada es la relación positiva entre la fracción de mutaciones beneficiosas y la adecuación biológica de la especie (Hietpas et al., 2013).

En el FGM existe lo que se conoce como un espacio fenotípico, un espacio n-dimensional donde cada eje corresponde a un rasgo de los organismos y cada punto en el espacio corresponde a un fenotipo con una combinación única de rasgos. Allen Orr (2000) identificó que el número de dimensiones en el espacio fenotípico es determinante en la definición de la DFE, encontrando que la probabilidad de que una mutación sea beneficiosa decae en función del número de dimensiones, conclusión que ha probado ser robusta a la modificación de los supuestos básicos de Orr (Welch and Waxman, 2003). Esta observación es llamada el costo de la complejidad, donde complejidad corresponde al número de dimensiones del espacio fenotípico, y sugiere que organismos simples van a experimentar una mayor cantidad de mutaciones beneficiosas que organismos complejos. Asimismo, se ha interpretado el costo de la complejidad como un costo en la tasa evolutiva, por ejemplo,

Haygood (2006) estudia el costo de la complejidad en estos términos, preguntándose por la correlación entre complejidad y tasa mutacional, donde el aumento de este último puede compensar el costo sobre la tasa evolutiva. Haygood observa que a medida que aumenta la complejidad de los organismos, medida como el número de tipos celulares distintos, aumenta el tamaño de sus genomas, esto tiene como consecuencia que ocurran más eventos mutacionales, aumentando la tasa mutacional, lo cual aumenta la tasa de fijación. Sin embargo, a medida que aumenta la complejidad, el tamaño efectivo de las poblaciones suele disminuir, lo cual tiene el efecto opuesto sobre la tasa mutacional, por lo tanto, el efecto final está mediado por el balance entre estos dos factores. De forma neta, los autores indican que hay una aceleración apreciable sobre la ocurrencia de mutaciones, sin embargo, esto parece no ser suficiente para compensar el efecto de la complejidad sobre la tasa evolutiva. Si bien, el efecto directo de la dimensionalidad sobre la tasa evolutiva se conoce en detalle, las posibles consecuencias del fenómeno del costo de la complejidad bajo distintas condiciones no han sido exploradas extensamente. Por ejemplo, Razeto-Barry et al. (2011) encontraron que el costo de la complejidad en un ambiente con cambios periódicos va a producir el efecto opuesto, donde los organismos más complejos acumulan mutaciones a una mayor tasa que los organismos más simples. Este hallazgo es interesante debido a que un ambiente variable es una suposición más realista al momento de evaluar este fenómeno en poblaciones naturales. En el FGM la variabilidad ambiental se modela por medio del cambio de posición del fenotipo óptimo y es equivalente a modelar procesos evolutivos de diferente extensión de tiempo, esta similitud se debe a que un ambiente muy variable corresponde a una sucesión de caminatas evolutivas de corta duración, mientras que un ambiente estable es equivalente a una caminata evolutiva extensa en términos del número de eventos mutacionales ocurridos.

Fisher (1930) describe que la adecuación de una población a una situación o a un ambiente puede ser entendida como la distancia entre dos puntos en un espacio,

donde el número de dimensiones corresponde al número de aspectos sobre los cuales la población puede estar más o menos adecuada al ambiente. En ese escenario, una mutación corresponde a un desplazamiento en ese espacio y, por lo tanto, a un cambio en el valor para cada uno de los aspectos. La propiedad de una mutación de tener efecto sobre varios rasgos fenotípicos de un organismo es llamada pleiotropía, consecuentemente, una interpretación biológica que se le da al número de dimensiones del espacio fenotípico de Fisher es el nivel pleiotrópico de una CDS (Gu, 2007). Interesantemente, al igual que el costo de la complejidad, se ha propuesto que la pleiotropía tiene el efecto de restringir la tasa evolutiva (Hodgkin, 1998), particularmente por medio de la pleiotropía antagónica, la cual hace referencia al fenómeno donde una mutación puede tener un efecto beneficioso en relación a un rasgo fenotípico, pero el efecto pleiotrópico sobre otro rasgo es deletéreo, generando un conflicto entre los efectos (Zhang, 2023). El costo de la complejidad puede ser entendido entonces como la formalización en el FGM de la relación negativa entre la pleiotropía y la tasa evolutiva. Estudios sobre la base molecular de la pleiotropía apuntan a múltiples mecanismos mediante los cuales una mutación puede modificar varios rasgos, por ejemplo, una mutación puede afectar un sitio regulador, como un enhancer, potencialmente modificando la expresión de varias CDS independientes (Zhang, 2023). Uno de estos mecanismos pareciera ser el más prominente y corresponde a la propiedad de un producto proteico de participar en múltiples procesos biológicos, mediado por el número de PPI y por el número de localizaciones celulares, pero no por el número de funciones moleculares (He and Zhang, 2006).

El efecto negativo que se propone para la pleiotropía es interesante porque tiene una relación directa con uno de los principios fundamentales de la evolución molecular, "*las moléculas o partes de una molécula funcionalmente menos importantes evolucionan (en términos de sustituciones mutacionales) más rápido que las más importantes*" (Kimura and Ohta, 1974). Como tal, ha habido múltiples intentos de ver si la predicción relativa

a la pleiotropía se cumple en sistemas biológicos reales (He and Zhang, 2006; Ericson et al., 2006; Salathé et al., 2006; Podder et al., 2009; Chakraborty and Ghosh, 2013; Pritykin et al., 2015; Chesmore et al., 2016; Chakraborty et al., 2016; Fraïsse et al., 2019; Rennison and Peichel, 2022; Williams et al., 2022). Sin embargo, los resultados no han sido concluyentes, por ejemplo, Fraisse y cols. descubrieron que, aunque los efectos negativos de la pleiotropía podían eludirse mediante cambios en la expresión génica, los valores intermedios de pleiotropía tendían a tener un impacto muy negativo en la respuesta a la selección natural direccional, mientras que Rennison y Peichel descubrieron que regiones de interés evolutivo vinculadas a adaptación reciente estaban enriquecidas en genes con pleiotropía intermedia. Este complejo escenario se ha profundizado aún más con la constatación de que, incluso en el modelo geométrico de Fisher, ha habido resultados contrastantes (Razeto-Barry et al., 2011; Razeto-Barry and Maldonado, 2011). En el presente proyecto de investigación proponemos profundizar nuestro conocimiento sobre el efecto que tiene la pleiotropía de las secuencias codificantes sobre sus tasas de acumulación de mutaciones. Esta profundización estará focalizada en un análisis temporal de las trayectorias evolutivas, intentando observar las distintas etapas de esta relación pleiotropía-tasa evolutiva y las condiciones bajo las cuales esta relación cambia, tanto a nivel de modelamiento de procesos evolutivos como a nivel de análisis de datos genómicos.

3.2 Hipótesis

El efecto de la pleiotropía sobre la tasa de evolución molecular de las secuencias codificantes es dinámico en el tiempo.

3.3 Predicciones

- El análisis temporal de una trayectoria evolutiva en el modelo geométrico de Fisher va a revelar un efecto dinámico de la dimensionalidad sobre la tasa evolutiva mediada por cambios en la contribución relativa de la selección natural y la deriva genética.

- La relación entre pleiotropía y tasa evolutiva de secuencias codificantes de pingüinos va a presentar un patrón temporal similar al efecto dinámico de la dimensionalidad sobre la tasa evolutiva descrito para el modelo geométrico de Fisher.

3.4 Objetivos

3.4.1 Objetivo general

Determinar la relación entre la pleiotropía y la tasa de evolución molecular a lo largo de una trayectoria evolutiva.

3.4.2 Objetivos específicos

1. Cuantificar la contribución relativa de la selección natural y la deriva genética en el modelo geométrico de Fisher.

2. Vincular la dimensionalidad con la contribución relativa de la selección natural y la deriva genética en una trayectoria evolutiva.

3. Analizar el cambio temporal del efecto de la dimensionalidad sobre la tasa evolutiva.

4. Traducir las variables del modelo geométrico de Fisher a medidores biológicos.

5. Evaluar la relación entre pleiotropía y tasa evolutiva en la historia evolutiva de los pingüinos.

3.5 Organización de la tesis

La presente tesis está estructurada entorno a la separación entre una aproximación de modelamiento y una aproximación bioinformática. El primer capítulo tiene tres secciones, donde la primera corresponde al análisis de difusión, que fue el método escogido para cumplir el objetivo específico 1. Aquí se puede encontrar el desarrollo de la ecuación diferencial parcial utilizada para modelar el FGM y la evaluación del ajuste entre la ecuación y las simulaciones. La segunda sección incorpora el primer y segundo objetivo específico haciendo utilización de una analogía a los sistemas de partículas estudiados en disciplinas como la mecánica de medios continuos. Finalmente, la tercera sección trabaja el tercer objetivo específico por medio de simulaciones y la utilización de las herramientas desarrolladas en las dos secciones anteriores. El segundo capítulo corresponde al análisis de secuencias codificantes de pingüinos y está dividido en dos secciones. La primera sección aborda la problemática de la relación entre pleiotropía y tasa evolutiva vinculada a los dos últimos objetivos específicos. La segunda sección explora la relación establecida entre la dinámica evolutiva encontrada para las secuencias codificantes de pingüinos y la dinámica evolutiva de las simulaciones en el FGM con el propósito de inferir información funcional relevante de la historia evolutiva de estos linajes. Hay una correspondencia directa entre los dos capítulos y la estructura del objetivo general, donde el primer capítulo se propone encontrar la forma específica en la que cambia la relación entre pleiotropía y tasa evolutiva en el FGM y encontrar la explicación mecánica detrás de esos cambios, mientras que el segundo capítulo aborda la comparación de los resultados *in silico* con los resultados en la historia evolutiva de los pingüinos.

4 Fisher's geometric model and the contribution of drift and selection on the dimensionality-evolutionary rate relation

4.1 Introduction

In this work, we focus on a particular type of molecular change that corresponds to the change in the sequence in which nucleotides are ordered along the DNA molecule. Mutations are one of the main sources of this type of variation, occurring first at the individual level, followed by a process of propagation within the population throughout the generations. This process depends on multiple factors, which can be broadly divided into two classes, selective and drift processes. Three possible outcomes can occur from the propagation process: fixation, loss, and maintenance of polymorphism, which correspond respectively to the replacement of the ancestral form by the mutant, elimination of the mutant and conservation of the ancestral form, and conservation of both in a simultaneous coexistence (Kimura, 1962).

While there are complex mechanisms by which the occurrence of mutations can be regulated, the mutational process depends mainly on cellular contingent events, therefore, at an evolutionary timescale, the mutation rate by individual by site is taken to be constant. After the first studies on molecular evolution on proteins, researchers were led to believe that, just like the mutation rate, the substitution rate was also constant for a given kind of protein, e.g., hemoglobin, irrespective of the lineage (Zuckerkandl and

Pauling, 1962; Margoliash, 1963), contrarily to what was expected from neo-Darwinian theory (Kimura, 1968; King and Jukes, 1969). A nourishing debate about the actual processes involved in molecular change through evolutionary time started, having as the main protagonist the neutral theory of molecular evolution (Kimura and Ohta, 1971; Kimura, 1983). This theory states that most of the mutations fixed are neutral concerning their fitness change effect and that their fixation is mainly due to genetic drift, thus the substitution process is roughly independent of the selective processes and as such, of the particularities of the "selection pressures" that each lineage experiences. As different proteins showed different rates of substitution, a new question arose, what properties that can be attributed to a given coding sequence or protein affect the rate of accumulation of mutations and in what way it affected it? Dickerson was one of the first researchers to pinpoint some of them, like the total surface area of a protein that is occupied in interaction with other molecules, crucial to its function, an example that Dickerson gives is the case of histone H4, which interacts strongly with other histones and is surrounded by the DNA strand, this degree of surface coverage is proposed to be partly responsible for its extremely low evolutionary rate. On the other hand, Dickerson suggests that the high evolutionary rate of molecules like fibrinopeptides and insulin peptide C is due to their high dispensability as, at the time, it was thought that these molecules were only collateral products of the processing of fibrin and insulin, respectively (Dickerson, 1971). Nowadays, we know the effects of many other variables, and most of them are explained by how they change the impact that new mutations have on the phenotype (Zhang and Yang, 2015; Alvarez-Ponce, 2020). On one side, variables that buffer the structural consequences of a sequence change have an accelerating effect, as is the case of the participation of chaperones in the folding of proteins (Bogumil and Dagan, 2010). On the other side, the variables that amplify the effects tend to increase the level of conservation of the sequences, as is the case of the expression level, where mutations have much bigger

effects upon misfolding because there are more products in the cell (Pál et al., 2001; Drummond et al., 2005).

The theoretical framework that conceptualizes the molecular evolutionary process as successive rounds of mutation and fixation (or loss) on a single locus dismissing stable polymorphisms has been called sequential fixation models, a type of origin-fixation model (McCandlish and Stoltzfus, 2014). Fisher's geometric model (FGM), defined briefly by Ronald Fisher in 1930, lies within this type of models, which has been used to answer a great variety of evolutionary questions concerning beneficial mutations (Hartl and Taubes, 1996), epistasis (Wilke and Adami, 2001; Martin et al., 2007; Gros et al., 2009; Hwang et al., 2017), dominance (Manna et al., 2011), fitness landscapes (Kopp and Hermisson, 2009; Gordo and Campos, 2013; Blanquart et al., 2014), recombination (Peck et al., 1997), molecular evolution theory (Razeto-Barry et al., 2012), selection pressure (Gros and Tenaillon, 2009), parallelisms (Chevin et al., 2010) and genotypic complexity (Lourenço et al., 2011), among many others (Tenaillon, 2014). In the present investigation we deal with an instance of this model with four major assumptions: the effective population size remains constant throughout the evolutionary process, phenotypic change in the population is due to the structural changes at a single locus corresponding to the coding sequence of a single product, the mutation rate at the target locus is slow enough such that there cannot be more than two allelic forms at a given time and, finally, we suppose the existence of an unimodal fitness landscape where its peak is referred to as optimum. Interestingly, under these assumptions, the selection coefficient distribution of the possible mutations is not predefined, as is the case for the house-of-cards (HOC) model (Kingman, 1977, 1978), but it emerges from the random phenotypic change caused by mutations and the fitness decay function.

Originally, Fisher (1930) used its model to make a point about the magnitude of phenotypic change of individual mutations and its place in meaningful evolutionary change.

He showed that, as the mutation magnitude increases, the probability of it being favorable sharply decreases. This line of inquiry was retaken by Motoo Kimura (1983), who showed that, while indeed small mutations share a higher probability of being favorable, they also have on average a smaller impact on the population fitness, resulting in the mutations of intermediate size of having the greatest probability of becoming fixed. Allen Orr (1998) subsequently enriched these results further by pointing out that the findings of Kimura were true only for the first mutation of an adaptive evolutionary process, but, as the fitness of the population changes by the successive fixation of mutations, the mean size of the mutations fixed decreases, shifting the curve found by Kimura toward smaller magnitudes.

Orr also implemented the FGM to research the effects of the number of phenotypic dimensions on the evolutionary process, and he found that a mutation of a given magnitude has a decreasing probability of being favorable as the dimensionality (number of dimensions) of the phenotypic space increases (Orr, 2000), a phenomenon known as the cost of complexity. In the same spirit as Orr (1998), we propose to extend these theoretical findings by following the changes in the effect of dimensionality throughout the evolutionary process. As ever-increasing degrees of organismal complexity are prevalent in evolutionary history, the cost of complexity poses an interesting challenge to evolutionary theory. Much has been studied about possible mechanisms in which this cost is bypassed (Welch and Waxman, 2003; Haygood, 2006; Wang et al., 2010; McGee et al., 2016) and how this dimensionality can be measured (Gu, 2007; Tenaillon et al., 2007). These studies relate directly to the study of pleiotropy and its effects on evolutionary dynamics, as pleiotropy is defined as the phenomenon where changes in a single coding sequence can potentially produce changes in multiple traits (Pavlicev and Cheverud, 2015; Zhang, 2023). The relation between pleiotropy and dimensionality is usually treated as follows, a complex organism has more traits that can change independently, so a given mutation experienced

by the organism has the potential to change more traits than mutations occurring in simpler organisms. If any change in a coding sequence can alter multiple independent traits is a different matter, most of the earliest models assumed universal pleiotropy, i.e., any coding sequence has access to change any trait in the organism, even so slightly (Zhang, 2023). Now, with more empirical data at hand, the main assumption is that of modular pleiotropy, where a given coding sequence has access to change only a subset of traits in the organisms (module) (Wagner et al., 2007), where the distribution of the number of traits affected by a coding sequence is L-shaped (Wang et al., 2010), i. e. many coding sequences can change a reduced number of traits, meanwhile, a small portion of them can change multiple independent traits. Most of the theoretical work done on the cost of complexity is based on the FGM, as its structure gives an intuitive way of modeling pleiotropy, where each independent trait is modeled as a single axis in phenotypic space, and modularity can be assessed by independent simulations.

The present extension on Orr's work is motivated by the discrepancy in empirical investigations on looking for a consistent negative effect of pleiotropy over the evolutionary rate, as most empirical evidence points to a mild negative effect or to no effect at all (He and Zhang, 2006; Ericson et al., 2006; Salathé et al., 2006; Podder et al., 2009; Chakraborty and Ghosh, 2013; Pritykin et al., 2015; Chesmore et al., 2016; Chakraborty et al., 2016; Fraïsse et al., 2019; Rennison and Peichel, 2022; Williams et al., 2022). Particularly illustrative is the case of the results of Chakraborty and Ghosh (2013) and Chakraborty et al. (2016), where the former reports a positive relationship between evolutionary rate and the number of associated biological processes of core and attachment proteins from protein complexes, meanwhile in the latter when looking at disease and non-disease associated genes, they found a significant negative correlation.

From a theoretical point of view, what should be expected about the effects of the evolutionary processes of selection and drift? Most molecular evolution in nature is sub-

jected simultaneously to processes related to natural selection and genetic drift, but their cause-effect relationships with variables like evolutionary rate or mutation magnitude, for example, are quite different. Discerning between both is not an easy task but is necessary to have a thorough description of the process. To address this issue, we again make use of a common tool in population genetics, diffusion analysis (Kimura, 1955, 1962; Ewens, 2004), which we apply to the FGM to create a method to identify the changes in the relative contribution of both processes. Here, diffusion analysis is meant to mean the description of the evolutionary process as modeled by the FGM, as the time change (or evolution in the mathematical sense) of the probability density function of the relative position of the population in phenotypic space with respect to the optimal phenotype, through a partial differential equation usually denoted as the Fokker-Planck equation or Kolmogorov forward equation (Ewens, 2004). The use of this tool is justified by the assumption that the nature of the FGM dynamics allows the evolutionary process to be approximated as a continuous stochastic process. We make use of the analogy with mechanical natural systems of particles moving in space, which provides us with intuitive notions to treat different types of movement and regimes. Unlike most other instances of application of this analysis in evolution, here we apply diffusion analysis not to the traditional propagation of an allelic form inside a population throughout the generations, but to the movement in phenotypic space of the population due to the successive fixation events. A similar approach can be found for example in the work of McCandlish et al. (2014) where the state variable corresponds to the fitness of the population, which can be directly translated to the state variable used in this work, the Euclidean distance between the population and the optimum phenotype. Nonetheless, a major difference between both approaches is that McCandlish et al. based their approach on the HOC model, whereas here we use directly the FGM in an origin-fixation fashion.

In the present chapter, we focus on the relationship between dimensionality and molec-

ular evolutionary rate from a theoretical viewpoint, using diffusion analysis and computer simulations to provide a better understanding of the changes in the relationship between both variables and the role of natural selection and genetic drift in those changes.

4.2 Materials and Methods

4.2.1 Fisher's geometric model

Simulations of the evolutionary trajectories of populations by the changes in a sequence using the FGM (Fisher, 1930) were implemented in Python with a custom script. The implementation proceeded as follows: a population's phenotype is defined by an array of n coordinates, where n is the number of dimensions of the phenotypic Cartesian space. Each time interval a random mutation is generated by creating an array of n coordinates, each drawn from a normal distribution of mean 0 and standard deviation 1 (Muller, 1959; Harman and V., 2010). This array is divided by its module to ensure that the magnitude of the mutation is 1. The actual magnitude of the mutation is drawn from a uniform distribution between 0 and 1 raised to the power of $1/n$ and multiplied by the maximum mutation magnitude. The power of $1/n$ is incorporated to ensure that each point in phenotypic space within reach of a mutation is equally likely to be selected, i.e. uniform density. The position of the mutant is then calculated as the displacement of the population's position by the mutation vector. The fitness of the mutants and the wild-type are calculated using a referential point in space, in this case, the origin of the phenotypic space, and a fitness landscape of Gaussian decay as the Euclidean distance to the optimum increases (Orr, 1998). Finally, we follow the derivation of Kimura (1962) to find the ultimate probability of fixation of the mutation (q) given the population size (N) and its selection coefficient (s). In Kimura's equation, the probability of fixation depends on the initial frequency and the time interval measured in generations, here we

are interested in calculating this probability for an initial frequency of 1/N (one individual of the species) and a time interval tending to infinity. This probability is calculated as

$$q = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}, \tag{4.1}$$

and, if this probability is greater than a random number drawn from a uniform distribution in the unit interval, then we assume that the mutation fixates and the population's position in phenotypic space is updated. This process is repeated until the target number of mutation events is achieved.

4.2.2 Fokker-Planck equation

Let us suppose that $f(x,t)$ is the probability density function that describes the probability of finding the population at a distance x from the referential point in phenotypic space with the maximum fitness (optimum) at a time t (measured in any evolutionary significant unit, as generations or million years, for example). Given the mutation rate μ we defined a time interval τ corresponding to the inverse of μ, which corresponds to the expected time it takes for a new mutation to arise. From the definition of τ we can define the probability that a population changes its distance to the optimum from $x - \zeta$ to x during the time interval τ. This probability ($p(x - \zeta, x)$) is equal to the probability that a mutation that generates a displacement of ζ in reference to the optimum occurs ($m(x - \zeta, x)$) and that such a mutation fixates ($q(x - \zeta, x)$).

The probability of finding the population at a distance x after a time interval τ is equal to the integral of the probability density function at position $x - \zeta$ times the probability of jumping to the position x over all possible values of ζ, i.e. all possible phenotypic

displacements within reach of a single mutation,

$$f(x, t+\tau)dx = dx \int_{-\infty}^{\infty} f(x-\zeta, t)m(x-\zeta, x)q(x-\zeta, x)d\zeta. \tag{4.2}$$

Equation 4.2 can be expressed as a partial differential equation by using the Taylor expansion around the point t for the function f approximated to the first derivative and the Taylor expansion around the point x for the function fmq approximated to the second derivative (see Appendix A.3),

$$\frac{\partial f}{\partial t} = \frac{1}{2\tau}\frac{\partial^2 f}{\partial x^2}\int_{-\infty}^{\infty} \zeta^2 m(x, x+\zeta)q(x, x+\zeta)d\zeta - \frac{1}{\tau}\frac{\partial f}{\partial x}\int_{-\infty}^{\infty} \zeta m(x, x+\zeta)q(x, x+\zeta)d\zeta. \tag{4.3}$$

The geometrical properties of the FGM were thus incorporated in equation 4.3 in the calculation of the probability of mutation. Every point inside the n-ball centered at the population's phenotype and with a radius equal to the maximum mutation magnitude is equally likely to be generated in a mutational event, as a consequence the probability of occurrence of any mutation of a given fitness effect is proportional to the area of the n-dimensional cap (at the surface of the n-ball centered at the optimum phenotype with a radius equal to the distance between the optimum and the mutant) formed by its intersection with the n-ball centered at the population's phenotype, where every phenotype in the surface of this n-ball has the same fitness (Fisher, 1930). When the cap area is normalized by the volume of the n-ball centered at the population's phenotype, it becomes equal to the probability that a mutation within the cap occurs during a mutational event. The specific cases and formulas used for the numerical simulations are specified in the Appendix A.2.

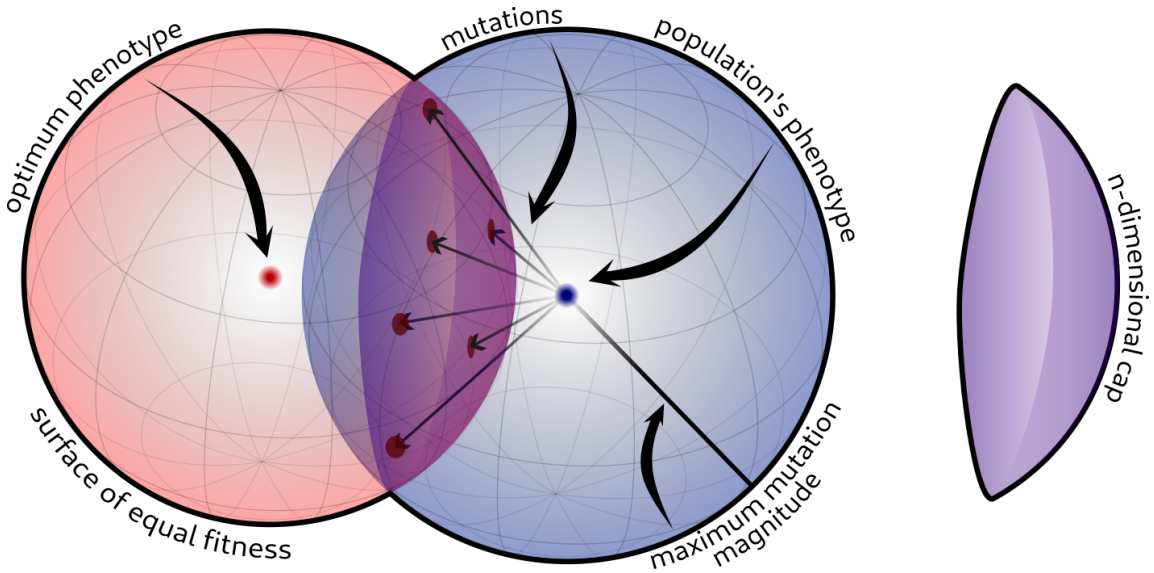

Figure 4.1: Probability of mutation. The probability that a mutant with a given fitness appears after a mutation event is equal to the surface of the n-dimensional cap belonging to the n-sphere centered at the optimum phenotype with radius given by the defined fitness (red sphere) formed by its intersection with the n-sphere centered at the population's phenotype with radius equal to the maximum mutation magnitude (blue sphere), divided by the volume of the n-sphere centered at the population's phenotype.

4.2.3 Numerical solution of the Fokker-Planck equation

To solve equation 4.3 the finite difference method was used (Arendt and Urban, 2020). This method requires the discretization of time and space to compute the probability density function as a system of linear equations. As the time is already discretized in the FGM, with intervals corresponding to mutational events, only the spatial variable was redefined as a finite number of equidistant points representing possible distances to the optimum. The partial derivatives of f were approximated by a Crank-Nicolson scheme (Crank and Nicolson, 1996), meanwhile, the two spatial extremes were treated as Neumann boundary conditions. The Neumann boundary condition defines that the

derivative of f at the border is equal to 0, this has the effect of a reflective border, which is accurate for a population that overshoots the optimum, as well as for a population infinitely far from the optimum where the probability of fixating a deleterious mutation tends to zero. The Neumann boundary conditions were implemented with backward and forward difference approximations at the near and far ends, respectively. With these specifications, the transition matrix was created. For every calculation a space bin length of 0.01 was chosen to avoid oscillations and the result was normalized to avoid mass loss in long calculations. The corresponding formulas are exposed in Appendix A.3.

4.2.4 Statistical analysis

To test the equation solution against the distribution of distances obtained from simulations in the FGM for a range of conditions we used the Kolmogorov-Smirnov test. The null hypothesis of the test is that the simulated distances could have been drawn from the distribution of the equation solution. On one hand, to research if the dimensionality affects the goodness of fit, we evaluated different populations in equilibrium (x_{eq}) as calculated by equation 4.4 (Tenaillon et al., 2007):

$$x_{eq} = \sqrt{-2 \ln \left(1 - \frac{1}{2N-1}\right)^{\frac{n}{2}}}. \tag{4.4}$$

On the other hand, we looked if, along an adaptive walk, there was a change in the goodness of fit by simulating multiple populations in a 2-dimensional phenotypic space starting at a distance 1.5 from the optimum. For each dimensionality condition a thousand simulations were performed (results shown in Figure 4.4 from Section 4.3.1).

4.2.5 Relative contribution of drift and selection

From the analysis of the evolutionary dynamics in the FGM with the Fokker-Planck equation, we derived a dimensionless number, known in the transport literature as the Péclet number (Pe), to estimate the relative contribution of selection and drift. The value assumed by the Pe given a set of conditions is a measure of the relative contribution of the first and second terms in equation 4.3, where the first term corresponds to the mean square displacement (MSD) associated with symmetric movement and diffusion, meanwhile, the second term corresponds to the mean displacement associated with asymmetric movement and advection. The Pe is a ratio which, when equal to 1, indicates that advection and diffusion are equally important to explain the dynamics of the system. Here we calculated the Pe as the mean displacement divided by the MSD of a population in a given position, this ratio is then divided by the characteristic length of the system, in this case, the maximum mutation magnitude was used (Gommes and Tharakan, 2020). Each plot from Subsection 4.3.2 was constructed using equation 4.3, either directly by evaluating the integrals as is done for Figures 4.5 and 4.6 (Subsection 4.2.2) or by solving the Fokker-Planck equation by the finite difference method (Subsection A.3.2).

The Pe was calculated for a thousand equidistant points between 0 and 1.5 distance from the optimum for each condition in Figure 4.7 to determine the distance in which the Pe is equal to 1. Additionally, the mean position of a population in a given time was used to calculate the time spent on a selection-driven regime. As the mean position is a function that decreases monotonically with time until it gets into the equilibrium, in a list of the mean positions in a trajectory of 10 million mutational events, a binary search algorithm was performed to find the point in time when the population crosses the Péclet number equal to 1.

4.2.6 Relation between dimensionality and evolutionary rate

The β index used in Subsection 4.3.3 corresponds to the slope of the linear regression between evolutionary rate and dimensionality. The molecular evolutionary rate was calculated differently depending on the figure, on one hand, for Figure 4.8 where it corresponds to the number of successful fixations over the number of total mutational events in a given time window. On the other hand, in Figures 4.9 and 4.10, the instantaneous evolutionary rate was calculated as the mean fixation probability over all possible mutations in a given point in phenotypic space.

4.3 Results

4.3.1 Diffusion analysis

The probability density function of the distance between the population and the optimum defined by equation 4.3 shows a monotonically decreasing mean distance that approaches asymptotically an equilibrium point. Meanwhile, starting from a distribution that is 0 everywhere except for the point at a distance of 1.5 from the optimum, as time progresses the variance of the distribution shifts from a sustained increase to a sustained decreased until it stabilizes at the equilibrium (Fig. 4.2). While no oscillations are observed in the distribution, a mild loss of mass has to be corrected by normalizing each final solution. This behavior mimics the bulk behavior of the FGM simulations as shown in Figure 4.3.

Based on the Kolmogorov-Smirnov test, the simulated data show a discordance with the equation's solution at the beginning and end of the adaptive walk for a population in a two-dimensional phenotypic space (Fig. 4.4.A). It was found that the expected theoretical distribution did not have a bad performance between 300 and 7000 mutational events, where the null hypothesis could not be discarded. Concerning the effect of dimensionality

in equilibrium, as expected from the performance of the equation's solution for a high number of mutational events, the theoretical distribution performed poorly, but with no appreciable effect of dimensionality on its performance (Fig. 4.4.B).

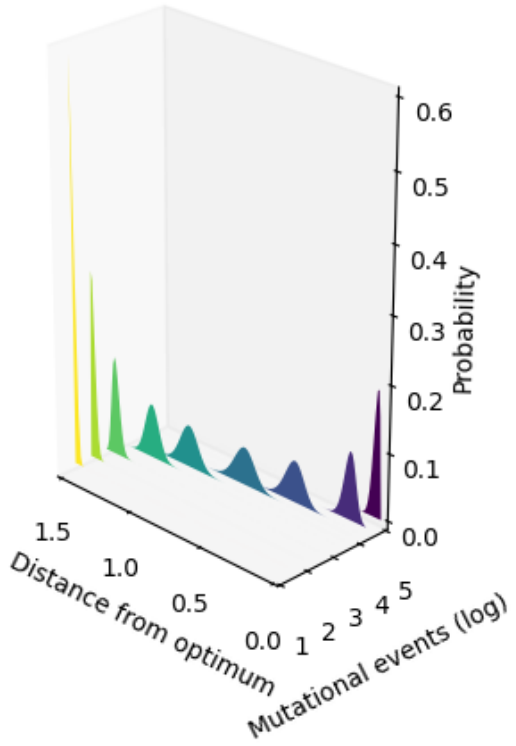

Figure 4.2: Fokker-Planck equation numerical solution for a population of 500 individuals starting at 1.5 from the optimum in a 2-dimensional phenotypic space and with a maximum mutation magnitude of 0.02. Each curve shows the probability distribution of the population at different times of the evolutionary trajectory.

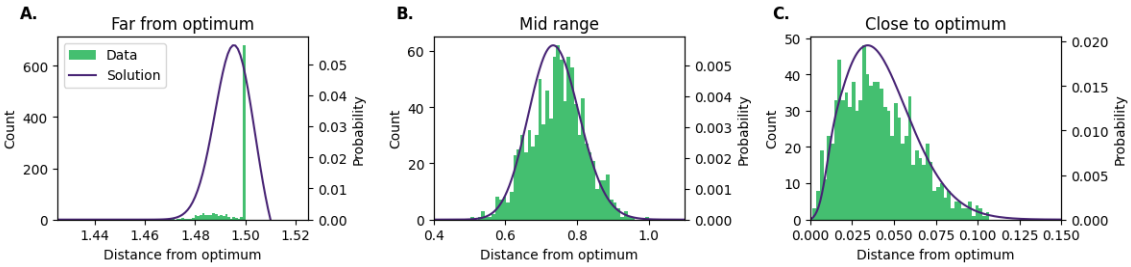

Figure 4.3: Exemplary comparison of the FGM simulations and the Fokker-Planck equation solution for a population with dimensionality 2, maximum mutation magnitude 0.02, and starting position 1.5 from the optimum. The histogram data for a thousand simulations is shown as green bars (count) and the discrete numerical solution for the Fokker-Planck equation for spatial discretization of size 0.001 is shown as purple lines (probability). From left to right the data is shown after 30, 7,196, and 100,000 mutational events.

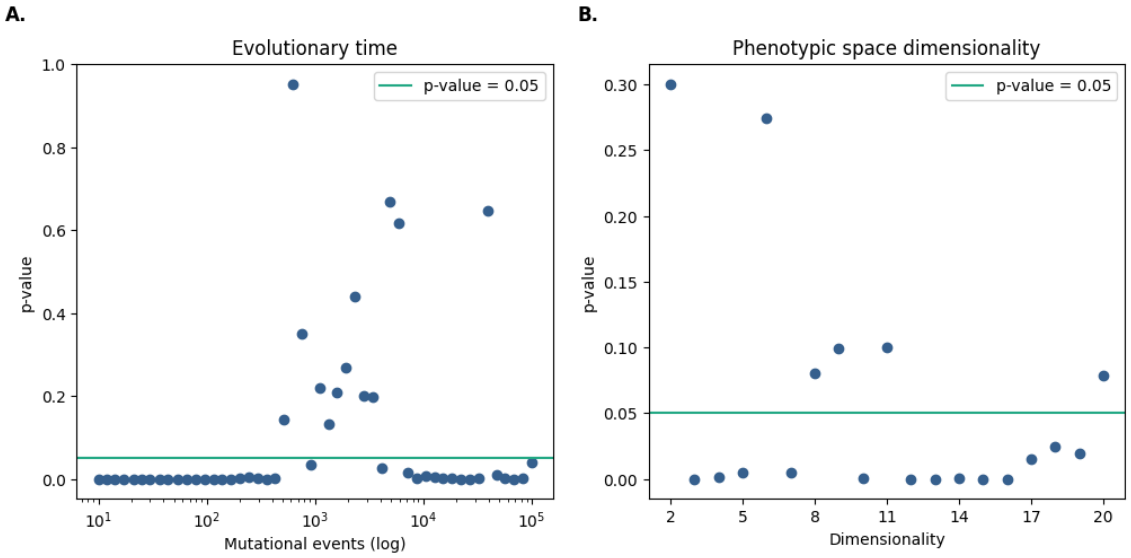

Figure 4.4: Kolmogorov-Smirnov test for the Fokker-Planck equation solution. A thousand FGM simulations with common parameters specified below were implemented for each dimensionality and the p-value of the Kolmogorov-Smirnov test was calculated. The common parameters of the simulations are a population size of 500 individuals, a maximum mutation magnitude of 0.02, (A) phenotypic space of 2 dimensions and a starting distance of 1.5 away from the optimum, and, (B) a starting distance corresponding to Tenaillon's equilibrium distance and a 100,000 mutational events. The horizontal line illustrates the significance level of 0.05.

4.3.2 Relative contribution of drift and selection

Three evolutionary regimes distinct with respect to the evolutionary forces present were tested by decomposing the Fokker-Planck equation in the mean square displacement (diffusion term) and the mean displacement (advective term). Genetic drift and natural selection were isolated by changing the probability of fixation formula. In Figure 4.5, populations with no drift (population size tends to infinity), shown as green curves, on average move towards the optimum along all the phenotypic space (positive values), with

a decrease in the velocity of movement as the distance from the optimum decreases. Any sort of movement reduces to zero at the position of the optimum. Populations with no natural selection (the selection coefficient of every mutation tends to zero), exhibited as purple curves, show movement as pictured by a not-null mean square displacement, but this movement has no specific direction far from the optimum. Close to the optimum, these populations show a strong movement away from it, reaching a maximum outward displacement of -3.6×10^{-5} when its position is exactly the optimum phenotype position. The populations under both forces (dark cyan curve) experience a shift in behavior from following the trend of the populations under selection exclusively, with an average movement towards the optimum, and adopting the behavior of escaping away from the optimum when the distance is smaller, like the drift-exclusive populations, reaching a mean displacement of -3.3×10^{-5} when the distance to the optimum is zero. In between both trends, an attractor can be found where the mean displacement of these populations is zero.

The Péclet number is calculated for the three regimes as a function of the distance to the optimum (Fig. 4.6). Natural selection-exclusive populations show a Péclet number always greater than 1, which spikes to greater positive values as the population approaches the optimum. On the other side, in drift-exclusive populations the Péclet number stays close to zero throughout most of the phenotypic space, this trend shifts close to the optimum at around a distance of 0.173, where the Péclet number crosses the critical value of -1 and stays lower it until it reaches the optimum at a zero distance. A minimum of -4.09 is observed around the distance of 0.02, which coincides with the maximum mutation magnitude. Again, the mixed regime shifts its behavior, having both last cases as upper and lower bound, respectively. The population reaches a Péclet number of 1 at around the distance 0.124 and crosses the critical value of -1 at around 0.073 from the optimum phenotype.

As critical evolutionary variables, we studied the effect of dimensionality and maximum mutation magnitude on two complementary indicators of selection and drift relative contribution to the evolutionary process. On one side, we estimated how much of the phenotypic space is dominated by selection and drift by the determination of the distance to the optimum where the Péclet number of a mixed population (selection + drift) is equal to 1 (Fig. 4.7.A), i. e. the point in phenotypic space where the contribution of selection and drift is equal. We found that the dimensionality has a weak positive relation with the distance of equal contribution, meaning that as the dimensionality increases, a bigger part of the phenotypic space is dominated by genetic drift. On the other hand, the maximum mutation magnitude does not have a linear relationship with the distance of equal contribution. For every dimension number, the minimum distance of equal contribution is achieved at maximum mutation magnitudes of around 0.06, which is approximately 0.146 with 20 dimensions and 0.044 with 2 dimensions. Selection is less dominant throughout the space for smaller and greater mutation magnitudes, more strongly with the former. On the other side, we calculated how many mutational events are needed for a population to cross the distance of equal contribution, starting from a place dominated by selection in an evolutionary trajectory (Fig. 4.7.B). Once again, there seems to be a direct effect of dimensionality, increasing the time spent in a selection-dominant zone, meanwhile, the maximum mutation magnitude minimizes the time spent in a selection-dominant zone at intermediate magnitudes, but this time, the minimum is achieved at around twice the maximum mutation magnitude encountered in the previous analysis, at a magnitude of approximately 0.115. Meanwhile, the decrease in mutation magnitude has a much less steep slope in time, for magnitudes over 0.32 there is a sharp increase in the time spent in the selection-driven zone, where the populations failed to cross the equal contribution distance in under 10 million mutational events. This threshold magnitude is increased with the decrease in dimensionality.

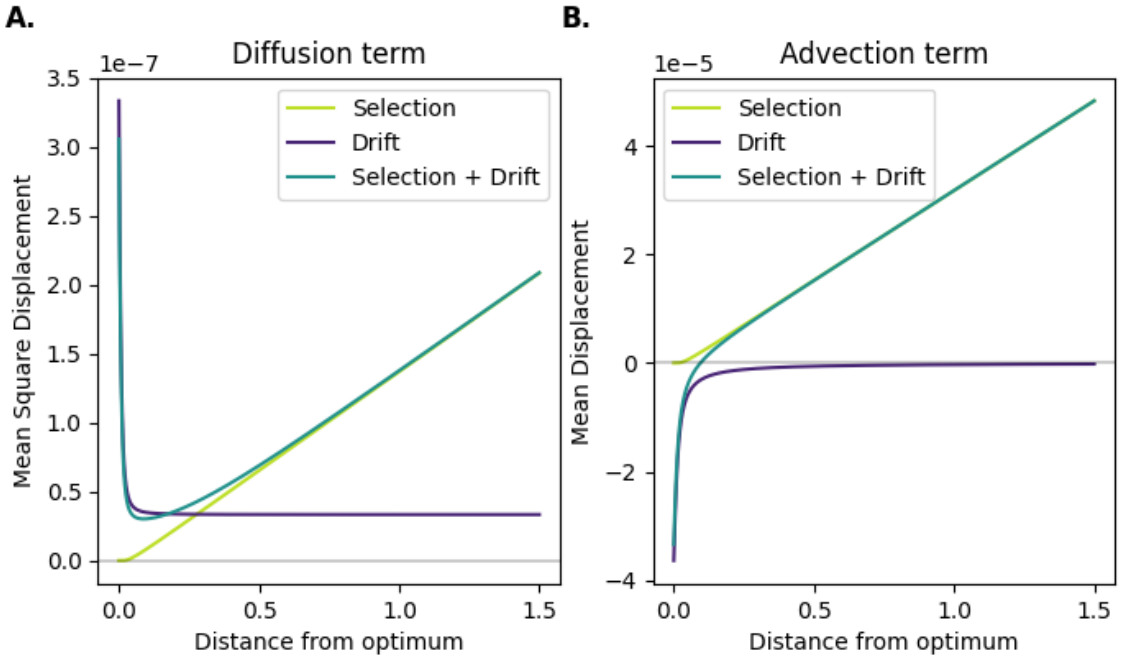

Figure 4.5: Theoretical movement of the populations in the phenotypic space as a function of the distance from the optimum. Mean squared displacement (A) associated with the diffusive movement and mean displacement (B) associated with the advective movement are calculated as the mean squared fixation magnitude and mean fixation magnitude, respectively. The displacement is calculated for three formulas of fixation probability; $1 - e^{-2s}$ for positive selection coefficients, 0 otherwise (selection), $1/N$ (drift), and $1 - e^{-2s}/1 - e^{-2Ns}$ (selection + drift). All calculations assume a maximum mutation magnitude of 0.02, a phenotypic space of 10 dimensions, and a population size of 500 individuals.

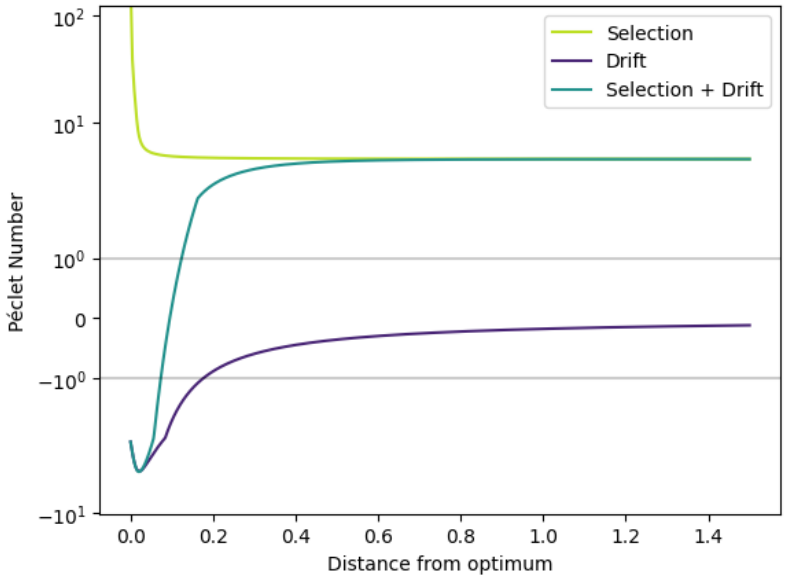

Figure 4.6: Péclet number in symmetric logarithmic scale as a function of the distance between the population and the optimum. The Péclet number is calculated for three formulas of fixation probability; $1 - e^{-2s}$ for positive selection coefficients, 0 otherwise (selection), $1/N$ (drift), and $1 - e^{-2s}/1 - e^{-2Ns}$ (selection + drift). All calculations assume a maximum mutation magnitude of 0.02, a phenotypic space of 10 dimensions, and a population size of 500. The diffusion, positive advection, and negative advection regimes are separated by horizontal lines at the critical values 1 and -1.

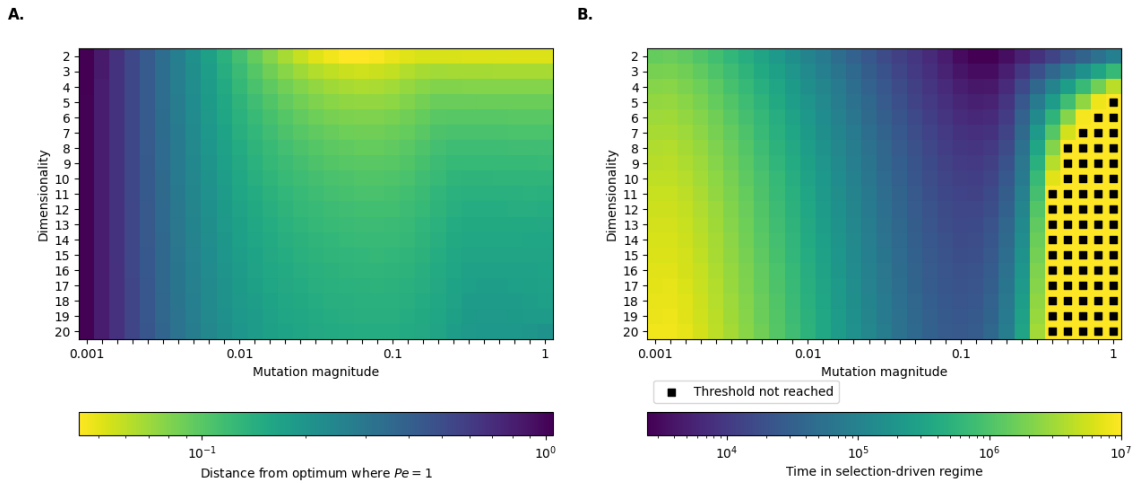

Figure 4.7: Effect of the dimensionality and the maximum mutation magnitude on the relative contribution of selection and drift in FGM trajectories with a population size of 500. (A) Heatmap showing the base 10 logarithmic distance from the optimum where the contribution of both forces is the same. Values closer to the optimum correspond to phenotypic spaces dominated by selection, and by drift otherwise. (B) Heatmap showing the base 10 logarithmic time that a population starting at a distance 1.5 from the optimum spends on average under a selection-driven regime. The black squares show the conditions under which the population does not change regime after 10 million mutational events.

4.3.3 Effect of dimensionality on the evolutionary rate

Three successive stages of dimensionality-evolutionary rate relation were found when simulating 950 populations in the FGM (Fig. 4.8). The evolutionary rate could not be calculated instantaneously but was approximated by taking the last 2,000 mutation events and counting the successful fixations during that period. A linear regression was performed for each time point to evaluate the association between dimensionality and evolutionary rate. As a result, we found that this relation changes dramatically at different

points in time from a significant negative relation to no relation at all, passing through a significant positive relation. The magnitudes of the evolutionary rates also change, being much greater at the start of the evolutionary process than at the subsequent stages.

Using the Fokker-Planck equation of the system, we calculated the differences in evolutionary rate between populations with phenotypic spaces of different dimensionality, summarized as the slope (β) of the linear regression between both variables, as a function of time. Every population parameter is the same except for the number of dimensions (Fig. 4.9). The same three stages were found, delimited by the change of sign of the slopes, with the first stage characterized by being brief, lasting less than 13,000 mutational events (Fig. 4.9) and by starting at strongly negative slopes (-0.00036). This stage is followed by a stage of positive slopes, which lasts 136,000 mutational events, approximately, and which reaches a maximum at around 30,000 mutational events. After this, the dimensionality-evolutionary rate relation decreases decelerating, in this final stage the slope value reaches a stable point at a value of -2.23×10^{-6}.

To have a detailed description of the changing relationship between dimensionality and evolutionary rate we inspected the change in the position of the populations relative to the optimal phenotype during an evolutionary walk (Fig. 4.10). The instantaneous evolutionary rate landscape in panel A shows the dependence of the evolutionary rate on the dimensionality as well as on the distance to the optimum, where the evolutionary rate increases sharply with this distance for lower dimensionalities. The iso-temporal cohorts of the 19 evolutionary walks, that started at a distance of 1.5 from the optimum, are not parallel but deflect towards the optimum for populations evolving in low dimensional phenotypic spaces. After 200,000 mutational events these differences in position are reduced, stabilizing with values ever-increasing with dimensionality, but nonetheless, very close to each other. Each cross-section represented by the isotemporal cohorts can be interpreted as an instantaneous picture of the velocity of the populations. Given that

in the figure time progresses from right to left, note that after the 10,000 mutational events, the initial trend of negative relation between dimensionality and evolutionary rate is reversed, from a difference given by an evolutionary rate of 0.0115 for the dimensionality of 2 and 0.0053 for the dimensionality of 20 at 1,000 mutational events to the difference given by an evolutionary rate of 0.00258 and an evolutionary rate of 0.00399, for the same populations, respectively. This difference is again reduced after 200,000 mutational events, where the evolutionary rate of the populations evolving in phenotypic spaces with dimensionalities 2 and 20 are 0.00193 and 0.00189, respectively. Meanwhile, panel B shows that the Péclet number has a stable value for a given evolutionary trajectory at the beginning of the process and drops significantly at later stages. Interestingly, the Péclet number is greater for populations evolving in high-dimensional phenotypic spaces.

At large distances from the optimum Figure 4.11 shows a negative relation between the Péclet number and evolutionary rate for a given point in phenotypic space, a trend that changes non-linearly as the distance to the optimum is reduced. For a given dimensionality, Figure 4.11 shows with no exception that the evolutionary rate spikes after a given Péclet number is reached, a value that is dimensionality-dependent.

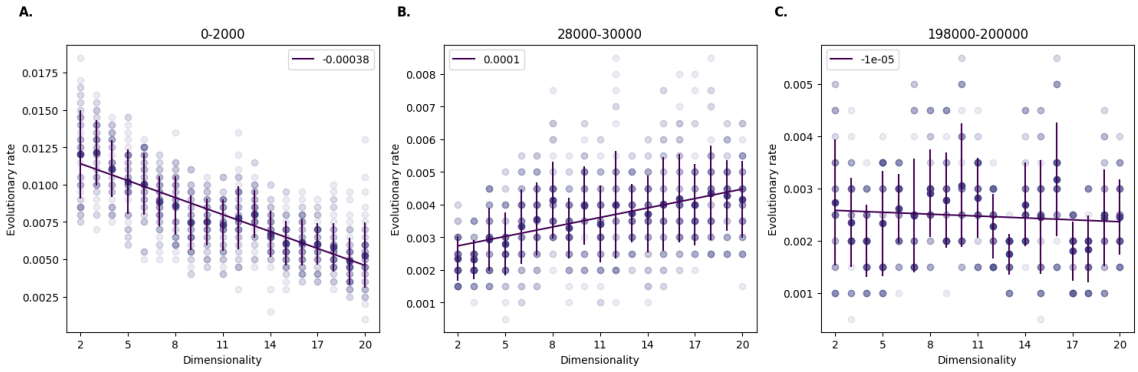

Figure 4.8: Evolutionary rate as a function of dimensionality on the FGM. 50 simulations for each dimensionality from 2 to 20 dimensions were carried out for 200,000 mutational events. The evolutionary rate is calculated as the proportion of mutations fixed during an evolutionary time interval of 2,000 mutational events. The result of each simulation is shown as transparent purple markers. From right to left the resulting evolutionary rates are shown for three time windows starting at 0, 28,000, and 198,000 mutational events. A linear regression is shown as a light purple line with the legend showing the slope. Vertical bars show the standard deviation from the mean evolutionary rate for each dimensionality.

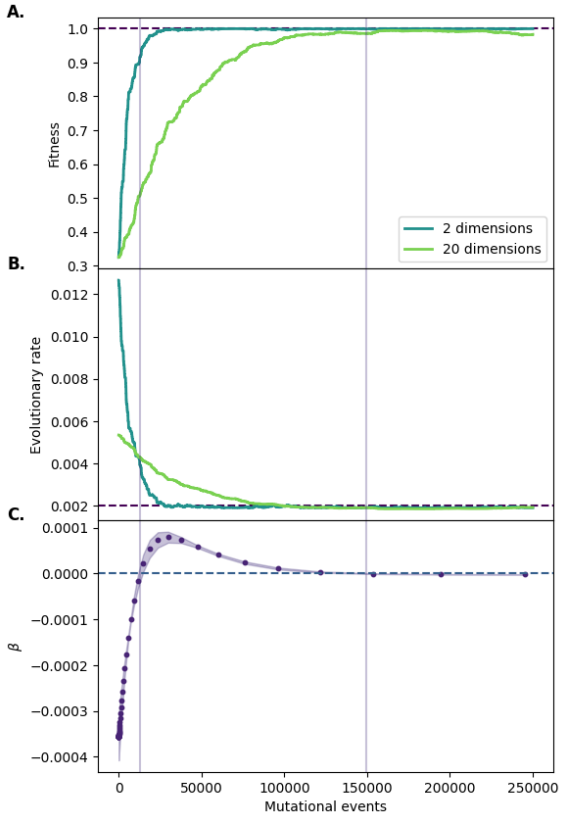

Figure 4.9: Relation between dimensionality and evolutionary rate as a function of evolutionary time. (A) Fitness of two populations simulated in the FGM as a function of evolutionary time. Both populations started at 1.5 from the optimum, with a maximum mutation magnitude of 0.02 and a population size of 500. Maximum fitness (1.0) is shown as a dashed line. (B) Evolutionary rate as a function of evolutionary time for the same populations of the top panel. The neutral fixation probability ($1/N$) is shown as a dashed line. (C) Slope of the linear regression (β) fitted to the data of evolutionary rate against dimensionality as a function of evolutionary time. The purple markers show the values of the slope at times spaced logarithmically. The standard deviation error for the slope derived from the regression is shown as a light purple shadow. The places where the slope is null are shown as vertical lines.

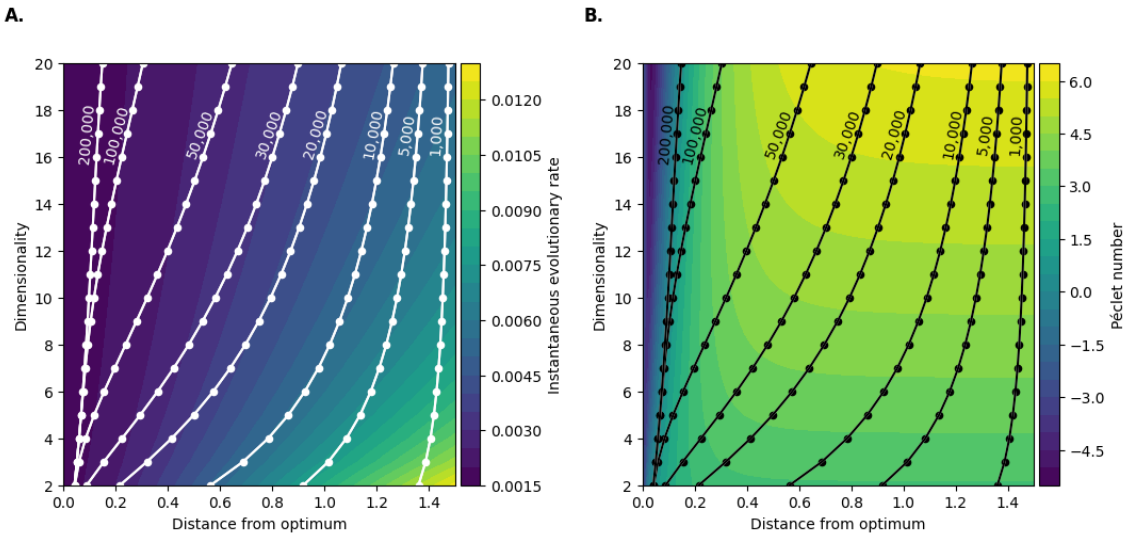

Figure 4.10: Relation between dimensionality, evolutionary rate and the Péclet number as a function of evolutionary time. (A) The heatmap shows the instantaneous evolutionary rate as a function of the dimensionality of the phenotypic space and the distance from the optimum. The white markers show the mean position of a population with a maximum mutation magnitude of 0.02, a population size of 500, and a starting position of 1.5 from the optimum at different moments in the evolutionary trajectory measured as the number of mutational events calculated with the Fokker-Planck equation. The same points in time for populations with different phenotypic space dimensionality are shown as white curves. (B) The heatmap shows the Péclet number as a function of dimensionality and the distance from the optimum. The black markers and curves are equivalent to the white markers and curves of A.

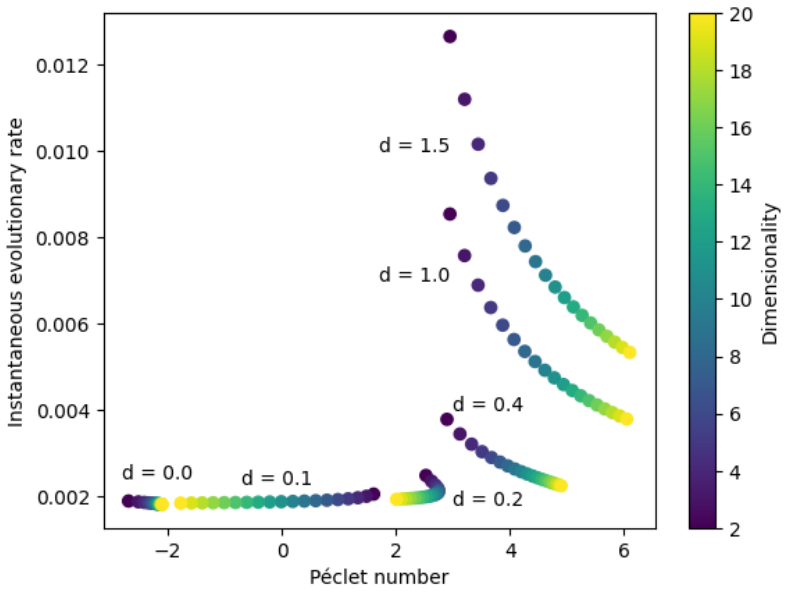

Figure 4.11: Relation between the Péclet number and the evolutionary rate. Each point corresponds to a single simulation static in a given place in the phenotypic space, with its dimensionality explicit in the color bar. There are six labeled groups of simulations, grouped by their distance to the optimum, where d corresponds to the optimum distance. Common population parameters correspond to a population of 500 individuals, and a maximum mutation magnitude of 0.02.

4.4 Discussion

The Fokker-Planck equation provides a method to describe deterministically the time evolution of the probability density function of the population's distance to the phenotypic optimum, avoiding the intricacies of handling the stochastic nature of the mutational processes (Arendt and Urban, 2020). Evolution under the geometric model of Fisher entails that no mutational event occurs when more than one allelic form is present in the population, this absence of overlap between mutation-fixation rounds ensures the Markovian nature of the system (Sella, 2009; McCandlish and Stoltzfus, 2014), where a

given state only depends on the previous one, a condition indispensable for a diffusion description. Even though the FGM system is continuous in (phenotypic) space, the same generation of bounded non-overlapping cycles of mutation-fixation provides that the system is discrete in time. This property may represent an unrealistic assumption when looking at real data, given the sheer number of bases in a single genome, nonetheless when considering only the phenotypic changes caused by mutations on one coding sequence, which can be around 1 kb, the assumption turns out to be more realistic. The discrete process is then approximated as a continuous stochastic process when doing the diffusion analysis, as the discrete character of the model is compensated by the large number of mutational events studied in these simulations.

Equation 4.3 has a general form that can describe a great family of systems, where the only component that is specific to the FGM is the mutation function $m(x, x + \zeta)$. The mutation function incorporates the geometrical considerations of the FGM, where its specific form is derived thanks to the assumption that any point in phenotypic space within the reach of a single mutation is equally likely to occur. This is not the case for most implementations of the FGM (Orr, 1998; Razeto-Barry et al., 2011; Ram and Hadany, 2015), as this assumption implies that the produced mutation magnitudes are not uniformly distributed, so when comparing between simulations with the same maximum mutation magnitude, but different dimensionality, the mean mutation magnitude increases with the number of dimensions. Nonetheless, this assumption simplifies the calculation of the mutation functions, as the fraction of possible mutant phenotypes is calculated with simple high dimensional geometric formulas (Li, 2011). To avoid misleading conclusions from simulations with this method of mutation generation, we repeated every simulation using the method specified by Ram and Hadany (2015). As a result, we were able to recover the three stages shown in Figure 4.9.C (data not shown), additionally it was noted that our assumption reduces the differences in evolutionary rates between

populations evolving in phenotypic spaces of different dimensionality, as a consequence, the dimensionality-evolutionary rate relation has the same sequence of sign change, but the magnitudes are much greater than the ones found with the method used originally in this work. Consequently, it is concluded that our results are robust to this assumption.

Figure 4.3 shows that phenomenologically, the diffusion approach and the FGM simulations have the same behavior. This is further supported by the Kolmogorov-Smirnov test (Fig. 4.4), which would have helped us discard this approach if the p-values stayed at values smaller than 0.05. The low performance at the beginning of the adaptive walk lasts around 300 to 400 mutational events, which corresponds to 3 to 4 fixation events. This discrepancy could be related to the method used in Markov Chain Monte Carlo (MCMC) runs known as the burn-in period, which corresponds to the discarding of the first n iterations of a Markov process (Johansen, 2010). This is done to find a starting point of high probability as, at the start of a run the system could be in a low probability state, thus not reflecting its most common behavior (which is a problem in short runs). This is the case of the FGM simulations, which start with an uncommon Dirac function-like distribution, and take several mutational events to reach the expected Gaussian-like distribution. We also were able to observe discrepancies in the state of equilibrium which starts after around 30,000 mutation events, nonetheless, the dimensionality was not a major cause of the solution's performance (Fig. 4.4.B). The discrepancy can be due to the method used to approximate the Fokker-Planck equation solution, the finite difference method, which requires a fine-tuning of the spatial discretization to have a good performance. This method was preferred over the search for an analytical solution to the Fokker-Planck equation because of the complexity of the resulting expression. The equation found corresponds to an advection-diffusion equation with variable coefficients, which means that not only the state variable (the probability density function of the phenotype position) vary spatially, but also the mean square displacement and the mean

displacement terms, corresponding in the transport phenomena literature to variable diffusion coefficient and flux velocity.

The isolation of evolutionary forces to investigate the specific effects of natural selection and genetic drift was achieved by changing the fixation function in our model, a function responsible for the only part of the process where these forces participate, as the mutation process is independent of selection and drift (Razeto-Barry and Vecchi, 2017). The classical fixation function derived by Kimura (1962) -by the use of diffusion analysis- depends only on two variables, population size, and the selection coefficient, as the third variable, initial allele frequency, is always the frequency of one individual in the whole population ($1/N$). The selection-exclusive fixation probability can be obtained at the limit of the population size going to infinity, a condition under which no genetic drift occurs, the expression obtained is the traditional fixation function used in most of the FGM simulations where drift is dismissed (Orr, 2000). This unrealistic situation is rapidly reached with slight increases in population size and is a good approximation.

The same procedure was performed to isolate the effects of genetic drift, where the selection coefficient for every mutation was supposed to be zero, turning the original function into the classical function of neutral evolution posed by Wright. Realistic conditions where this could happen in natural populations can occur, like with very flat fitness landscapes, or what happens with most synonymous mutations, where the structural change of the genetic molecules has no appreciable effect on the protein structures.

Equation 4.3 can be divided into two components, the advective and diffusive terms. The first one is the sum of every mutation effect on optimum distance, multiplied by its probability of mutation and fixation, which is equivalent to the mean change in optimum distance experienced by a population at a certain point in phenotypic space. As this corresponds to a sum, every two opposing movements with the same probability of occurring would nullify, effectively eliminating every symmetric displacement of the population with

respect to the optimum, preserving only asymmetric movement. On the other side, the diffusive component is the same as the first one, but all displacements are squared, so opposing movements now would add up instead of canceling. This calculation is equivalent to the calculation of the mean squared displacement (MSD), which is usually used in the study of diffusion phenomena as a measure of random movement. The property of incorporating asymmetric as well as symmetric movement means the MSD could be non-zero even in the absence of diffusive processes.

In Figure 4.5 the selection-exclusive populations show positive mean displacement and MSD for most of the phenotypic space. A non-negative mean displacement means that populations evolving only under natural selection always move primarily toward the optimum. This makes sense as we know that the fixation probability of any mutation that decreases the fitness of the population is zero, so every possible displacement is towards the optimum. This also implies that the registered non-zero MSD is exclusively a consequence of this asymmetric and directed movement, as we can discard any symmetric movement from the very definition of the selective processes. Consequently, we can attribute nothing of the diffusive component to the selective force. As the population is closer to the optimum the average displacement decreases linearly, this is a consequence of the nature of the selection coefficient calculation, which depends on the ratio between phenotypes' fitness, added to the fact that the mutation magnitude is not a function of the distance to the optimum. This means that, as the population is closer to the optimum, the difference in fitness between the wild-type and the mutant is roughly constant, but the total value of the wild-type and mutant fitness increases, decreasing the value of the ratio. Accompanying this effect of distance, we know that the fraction of beneficial mutations also decreases, from the limit of 50% very far from the optimum, to 0% at the optimum as every possible mutation that changes the phenotype of the population has a deleterious effect.

On the other side, when analyzing the behavior of the drift-exclusive populations (Fig. 4.5) we also find non-zero (and nearly constant at a value approaching 3.334×10^{-8}) MSD for most of the phenotypic space. We can discard that this MSD is due to asymmetric movement because, throughout this section of the phenotypic space, the mean displacement of these populations is about -2×10^{-7} (away from the optimum), which corresponds to an MSD of 4×10^{-14}, where the registered MSD is six orders of magnitude greater than this. As the population approaches the optimum, a strong movement away from the optimum is shown. As the fixation probability of every mutation under drift-exclusive conditions is the same, the only explanation for the asymmetric movement to occur is in the mutation processes, which presents a bias away from the optimum which increases as the distance to the optimum decreases, for the reasons exposed in the previous paragraph. In summary, accounting for the behavior of both types of populations studied, we can affirm confidently that any positive mean displacement can be attributed exclusively to the action of natural selection, meanwhile, every diffusive movement, as well as a mean negative displacement can be attributed exclusively to the action of genetic drift.

The population which experiences the action of both forces simultaneously has an interesting behavior as changes its dynamics as a function of optimum distance, resembling both previously discussed populations. We can see that in the transition between one behavior to the other, the population crosses a position where the mean displacement is zero, which in the case of Figure 4.5 occurs at a distance from the optimum of around 0.095. Interestingly, if the position of the population changes from this point away from the optimum, the mean displacement turns towards the optimum, and the exact opposite occurs if the population moves from this zero point towards the optimum, moving away from it on average. This means that we are looking at an attractor to where the population is drawn, and which corresponds to the equilibrium distance, and consequently, the

equilibrium fitness, known to not be equal to 1 due to the genetic load (Poon and Otto, 2000).

At transport phenomena research, we find the utility of dimensionless analysis to correctly approximate difficult state equations that describe the complex dynamics of the target systems. Avoiding the actual physical units of the variables at play by scaling them, allows easier comparison between the terms of an equation and the subsequent neglect of the smaller ones, in favor of having a simpler equation to solve (Rapp, 2017). In the study of diffusion and matter transport, the traditional diffusion-advection equation, equivalent to our Fokker-Planck equation, is converted into a dimensionless equation by scaling the variables with a characteristic length, which in this case is taken to be the maximum mutation magnitude. In the dimensional analysis of this equation, the Péclet number (Pe) arises as the dimensionless number that weights the contribution of each term, indicating which one can be neglected for the approximate solution to retain its explanatory power of the system's behavior. Great values of Pe mean that the diffusive component can be neglected from the equation, and the solution can still be a good description of the system's dynamics, the opposite goes for Pe values close to zero. Analogously, using this property of Pe, we use it as an indicator of the contribution of each force to the system's behavior, where great positive values of Pe mean that we can neglect the contribution of drift, and the system closely resembles a selection-exclusive population. Meanwhile, values of Pe close to 0 or great negative values suggest that we can neglect the selective processes, in favor of explaining the populations' dynamic under drift-exclusive conditions. We found that the Pe changes as a function of the distance to the optimum, where greater distances favor values greater than 1, and smaller distances favor Pe values less than 1. As expected, the distance at which the critical value of $\mathrm{Pe} = 1$ is achieved moves as a function of different evolutionary variables, for example, as the population size decreases, this threshold moves further away from the optimum, meaning

a more prevalent contribution of genetic drift to the evolutionary process throughout the phenotypic space (data not shown).

We explored the change in the relative contribution of both evolutionary forces related to the Péclet number as a function of two evolutionary variables of particular relevance, dimensionality, and maximum mutation magnitude. This was done with respect to two complementary approaches, a spatial and a temporal approach in the context of an adaptive walk (Orr, 1998). Figure 4.7.A answers the question of how much of the phenotypic space is dominated by drift and how much is dominated by selection, while Figure 4.7.B answers the question of how much time a population in an evolutionary trajectory spends in a selection-dominated zone of the phenotypic space, and how much time in a drift-dominated zone. In general, both patterns seem to be opposite, while most of the phenotypic space is drift-dominated in trajectories with high dimensionality and low mutation magnitude, populations evolving under those conditions take the longest to enter the drift-dominated zone (except for high dimensionality and high mutation magnitude). The exact opposite can be said of spaces with low dimensionality and intermediate mutation magnitude, where most of the phenotypic space is selection-dominated, and the populations rapidly enter the drift-dominated zone. The reason behind these results is related to the famous analysis of Kimura over the magnitude of the mutations fixed in evolution. He corrected the argument of Fisher in favor of micromutationism by indicating that the magnitude of a mutation not only changes its probability of being favorable, as we have seen already but also decreases the absolute value of its selection coefficient (Orr, 1998). Therefore, for a favorable mutation, a smaller mutation magnitude also implies a smaller selection coefficient, and so, a smaller probability of fixation. So there exists a range of intermediate mutation magnitudes where both effects are balanced, where the probability of fixation is maximal, and consequently, the evolutionary rate is maximal. This same reasoning applies to the relative contribution of selection and drift, where the

main variable that changes due to the change in mutation magnitude is the selection coefficient (it does not affect population size), which is central to the balance between selection and drift. So, the regions dominated by selection are also the regions with the fastest evolving populations, providing that these populations transition faster from the starting selection-dominated zone into the drift-dominated equilibrium zone.

From this analysis, one can say that natural selection accelerates the accumulation of mutations over the basal rate provided by neutral processes. Where is the effect of purifying natural selection which is supposed to have a conservative effect causing the slow-down of the evolutionary rate? Natural selection is going to have a negative net effect on the evolutionary rate when most of the mutations are deleterious, which happens close to the optimum, but as we discussed already, close to the optimum the mutant and wild-type fitness are very high, causing a drastic decrease in the magnitude of the selection coefficients, decreasing not only the number of beneficial fixations favored by directional natural selection but also decreasing the deleterious fixations disfavored by purifying natural selection. Consequently, in the equilibrium, close to the optimum, the rate of evolution very closely resembles the neutral rate of evolution, as can be noted in Figure 4.9.B, where the probability of fixation in equilibrium is approximately 0.002, the reciprocal of their population size (500 individuals). Then, it is to be expected that smaller mutations have lower selection coefficients, so the evolutionary rate decreases, but mutations too large are very rarely beneficial, so the fixation rate is also small, as is shown by the black squares in Figure 4.7, where the populations take at least 10 million mutational events to cross the Péclet threshold.

Here we tried to contribute to the general discussion in evolutionary biology about the opposition between natural selection and genetic drift, nonetheless, the restrictions of the model do not allow us to explore the effects of neutral networks (Manrubia and Cuesta, 2015), constructive neutral evolution (Brunet and Doolittle, 2018) and neutral

exploration of rugged and multimodal fitness landscapes. In the same vein, the prohibition of polymorphisms in the model causes us to be unable to investigate the effects of the standing genetic variation and the facilitation of mutation combinations in the genesis of biological traits through natural selection (Beatty, 2016) against the effects of macromutations (Orr and Coyne, 1992), which poses the question about the creative nature of selection in the evolutionary process (Razeto-Barry and Frick, 2011).

As shown in Figure 4.8, the evident negative effect of dimensionality over the pace of substitution accumulation is not sustained during an evolutionary bout of 200,000 mutations. When looking at three particular points in the evolutionary history of 950 simulated populations, three radically different associations between dimensionality and evolutionary rate are revealed, where only the first one shows the usually expected negative effect of dimensionality see Section 4.1). From the start of the evolutionary walk, we can see that, even for remarkably favorable conditions to fixate mutations, only around 1% of them are fixed in the most rapidly evolving populations, and this rate only decreases as time progresses. This low evolutionary rate makes it difficult to have good statistics about the behavior of these simulated populations, as a low number of fixations may undermine an accurate representation of their evolutionary rate. For this reason, we chose a big window of time to record the fixation events, with the disadvantage of losing temporal resolution of the relationship between dimensionality and evolutionary rate. This approach resembles the real constraint of estimating the evolutionary rate of real coding sequences, where the instantaneous evolutionary rate is not known, but must be estimated through the changes recorded during a given interval of time, which, in most cases, implies having a mean evolutionary rate between different stages of dimensionality-evolutionary rate relation.

We used the Fokker-Planck equation to visualize the changing relation between dimensionality and evolutionary rate in a greater temporal resolution using the mean fixation

probability over all mutations possible for a given population's phenotype as an equivalent measure of instantaneous evolutionary rate. For a period of 250,000 mutational events, we used the slope of the linear regressions (β) between dimensionality and instantaneous evolutionary rate as an index of the dimensionality effect, even though we have no reason to expect a linear relation over any other kind of relationship defined by a more complex function (see Razeto-Barry and Maldonado (2011)). As shown by Figure 4.9, we have a continuous change of β related to the velocities by which the example evolving populations reach their corresponding equilibrium states. Here we can see the sustained increase in fitness of both populations due to the fixation of primarily beneficial mutations, which reduces the distance between population and optimal phenotype. Even though both populations start at the same distance from the optimum, shown by the same initial fitness, populations evolving in low-dimensional phenotypic spaces have a much greater initial evolutionary rate than the other populations, a difference responsible for the initial slope of -0.00036. This trend immediately starts to decrease until reaching a maximum after changing sign, which can be associated with the arrival of the populations evolving in low dimensional spaces to their respective equilibriums, as the exemplary simulation with two dimensions in igure 4.9. After that happens, the stability of the simulations with low dimensional spaces causes the sustained decrease in the difference between populations, as every population is on average decreasing its distances to the optimum and decreasing its evolutionary rate. This latter point is again a consequence of the decrease in the proportion of beneficial mutations among all possible mutations and a consequence of the decrease in selection coefficient magnitudes due to the increase in fitness, making the effect of selective processes less prevalent, in favor of a major contribution of the neutral fixation of mutations. Interestingly, we do not have a null relationship at the third stage, but a significant, though considerably small negative association between pleiotropy and evolutionary rate. Even though the magnitude of this relation is not comparable to the

magnitudes of the minimum and maximum of the first and second stages, respectively, it is consistent in time. This could be due to the imbalance of two properties of the FGM as dimensionality modifies the proportion of beneficial and deleterious mutations, increasing the genetic load for higher dimensions, decreasing the equilibrium fitness, and at the same time decreasing the instantaneous evolutionary rate. Both properties have opposite effects, so the results suggest that the latter property has a stronger contribution, generating a mild negative effect of dimensionality over the evolutionary rate at equilibrium.

To have a clearer view of the mechanics behind this phenomenon, we calculated the mean phenotypic position as a function of time for each evolutionary walk with different dimensionalities using the Fokker-Planck equation and contrasted them against the theoretical instantaneous evolutionary rate and the Péclet number. In Figure 4.10, every population starts at a distance of 1.5 from the optimum, and after 1,000 mutational events, we recorded their average position relative to the optimum. We can see that even after only 1000 mutational events, which only corresponds to around 7 to 12 substitutions, the simulations in phenotypic spaces with low dimensionality are already ahead of the other simulations. This difference only increases with time as shown by the subsequent snapshots of their trajectories from 5,000 to 30,000 mutational events. As this happens, their instantaneous evolutionary rate drops crossing to the second stage described previously. Interestingly, during the evolutionary bout, these results show that there is a point around the 10,000 mutational events where an intermediate number of dimensions, between 4 and 12, are the fastest evolving populations, fact that was obscured in previous analysis looking only at the linear regressions, and making evident that we are not to expect a linear relationship. It is evident from Figure 4.10, particularly in low dimensional trajectories, that there is a point where the mean stops moving, this happens after around 50,000 mutational events for populations with 2 and 3 dimensions

but takes much longer for higher dimensional spaces.

Concerning the Péclet number, Figure 4.10.B shows that a higher Pe does not imply a higher evolutionary rate, which could be the expected result since the evolutionary rate decreases along an evolutionary trajectory because the contribution of natural selection decreases. From Figure 4.11 we can see that far from the optimum, a bigger Pe is associated with a lower evolutionary rate when the dimensionality varies. As the dimensionality does not affect the genetic drift's effect on the probability of fixation, the only possible explanation is the role of natural selection in the decrease in evolutionary rate. This suggests that at the start of an evolutionary trajectory, the negative or purifying natural selection is going to play a key role in the reduction of the evolutionary rate exhibited by populations evolving in high dimensional phenotypic spaces. As we have seen, Fisher's geometric model predicts a dynamic relationship between dimensionality and evolutionary rate dependent on the number of mutations fixed and the distance between the population's phenotype and the optimum phenotype. This dependency is captured by the concept of time when the different evolutionary trajectories start at the same distance to the optimum. Far from the optimum the Péclet number indicates a negligible contribution of genetic drift, as a consequence, after a sudden change in the optimum's position or the population's phenotype, the evolutionary trajectory of the population starts at a zone in phenotypic space dominated by natural selection. However, as shown in Figure 4.11, this contribution of natural selection constrains the rapid fixation of mutations in high dimensional phenotypic spaces and facilitates that populations in low dimensional spaces spend less time in selection-dominated regimes, entering sooner to the drift-dominated zone, which has an evolutionary rate close to the neutral expectation. As a result, a changing pattern of the dimensionality-evolutionary rate relation emerges, which was divided into three major stages in this research, a strongly negative relationship, a long and positive relationship, and an almost null relationship.

5 Empirical insights on the pleiotropy-evolutionary rate relation

5.1 Introduction

Evolutionary biology has a rich collection of mathematical and computational tools used from population genetics to phylogenetic inference. Within these tools, we find a great body of research concerning the modeling of evolutionary processes spanning a wide variety of phenomena, a method of investigation that has been increasing since the 90s (Morozov, 2013). Ronald Fisher, whose work was seminal for the development of population genetics, devised a simple model that bridged genetic and phenotypic changes (Fisher, 1930) which has been used more and more often in the evolutionary literature (Tenaillon, 2014). Fisher's geometric model (FGM) has four major components: (i) it conceptualizes the phenotype as a collection of independent trait values, where the set of all possible phenotypes forms a phenotypic space, with a dimensionality equal to the number of traits, (ii) changes in the position of a population in this space are caused by mutations fixed in the population, therefore, such mutations correspond to vectors of displacement with direction (what trait(s) does the mutation affect) and magnitude (how much does the trait(s) change), the generation of such vectors must follow some rules related to the random nature of the mutation process, for example, the direction must be a random variable uniformly distributed, (iii) it assigns every point in phenotypic space a fitness value (a positive real number) calculated with two objects, a reference point which has the maximum fitness value and is considered the optimum phenotype, and a fitness

decay function, which typically has as input the euclidean distance between the population and reference points and, finally, (iv) it controls the movement of the population in space by the acceptance or rejection of displacement vectors by the calculation of a fixation probability which typically depends on the population size, a positive integer assigned to the population, and on the selection coefficient of the displacement vector, which depends of the fitness value of the points before and after the displacement. The strong simplifications assumed by the FGM (Martin and Lenormand, 2006b) coupled with its easily interpretable and robust predictions have led to multiple test attempts in real-world biological systems (Burch and Chao, 1999; Martin and Lenormand, 2006a; Velenich and Gore, 2013; Weinreich and Knies, 2013; Perfeito et al., 2014; Blanquart and Bataillon, 2016; Moutinho et al., 2022).

One of its major predictions concerns what Allen H. Orr coined as the cost of complexity (Orr, 2000). He noticed that, as the number of dimensions in phenotypic space increases, the fraction of randomly generated vectors of a given size that could reduce the distance between the population and reference points decreases. This means that, as we encounter organisms with an increasing number of traits, such as the difference between a single-cell bacteria and a multicellular chordate, it should be less likely to encounter a mutant individual with greater fitness than the wild-type population. An alternative interpretation can be made by equating the dimensionality with the number of independent traits that a mutation can potentially change, which is determined by the location of the mutation, i.e. the particular genetic structure where it occurs. If the structure corresponds to a protein-coding sequence, then the number of independent traits potentially affected by a given mutation on that sequence is defined by its pleiotropy (Gu, 2007; Razeto-Barry et al., 2011). On the other hand, the probability of a new mutation of a given size affecting the phenotype of being favorable impacts directly on the evolutionary rate of molecular change, as this rate is defined as the proportion of mutations

that get fixed, fewer mutations being favorable means that the probability of fixation decreases (Orr, 1998). Therefore, another interpretation of the cost of complexity is that more pleiotropic sequences are going to evolve (accumulate mutations) at a lower rate than less pleiotropic sequences. This interpretation is interesting because it has a direct relation to one of the founding principles of molecular evolution, "*functionally less important molecules or parts of a molecule evolve (in terms of mutant substitutions) faster than more important ones*" (Kimura and Ohta, 1974). As such, there have been multiple attempts to see if the prediction holds in real biological systems (He and Zhang, 2006; Ericson et al., 2006; Salathé et al., 2006; Podder et al., 2009; Pritykin et al., 2015; Chesmore et al., 2016; Chakraborty et al., 2016; Fraïsse et al., 2019; Rennison and Peichel, 2022; Williams et al., 2022). Although, most of them tend to show evidence that the cost of complexity is effective for pleiotropy measurements, usually the conclusions need to be nuanced in different directions, for example, Fraïsse et al. found that although the negative effects of pleiotropy could be circumvented by changes in gene expression, intermediate values of pleiotropy tended to have a much negative impact on the response to directional natural selection, meanwhile Rennison and Peichel found that regions of evolutionary interest concerning recent adaptation where enriched in genes with intermediate pleiotropy. This complex setting has been deepened further by the realization that under high environmental variability, the number of dimensions in Fisher's phenotypic space increases the rate of substitution (Razeto-Barry et al., 2011; Razeto-Barry and Maldonado, 2011), finding that has also received empirical support (Chakraborty and Ghosh, 2013).

The availability of genomic data has instantiated a new era in population genetics and evolutionary biology in general (Casillas and Barbadilla, 2017), as now we can test the vast theoretical literature on model and non-model organisms, with varying degrees of diversity, within-population genetic data and eco-evolutionary history information. One

such non-model group with a rich evolutionary history is penguins, grouped under a monophyletic clade known as Spheniscidae, comprehending almost 20 extant species. The last common ancestor of the crown group is estimated to have lived during the early Miocene 21.9 Mya in a sub-Antarctic environment (Vianna et al., 2020), radiating afterward to Antarctic and temperate zones of the Southern Hemisphere, with even a species crossing into equatorial waters, the Galapagos penguin. This biogeographic range makes the evolutionary history of penguins marked by the cycles of glaciation, as well as the particularities of the Antarctic and sub-Antarctic currents and fronts (Vianna et al., 2020). Added to the fact that penguins are the only extant birds that have lost completely their ability to fly and have made diving in oceanic waters a constitutive part of their way of life, makes them an interesting group for evolutionary studies (Frugone et al., 2019; Pertierra et al., 2020; Cole et al., 2020; Pirri et al., 2022).

Here we test the hypothesis of the cost of complexity under the evidence that the pleiotropic effect changes as a function of evolutionary time (see Chapter 4) using genomic data of 15 penguin species and, given the establishment of a relation between their evolution and the FGM, deduce relevant features about their history of diversification and change.

5.2 Materials and methods

5.2.1 Genomic data

Extracted coding sequences (CDS) from the genomes of 15 penguin lineages performed by Vianna et al. (2020), were used for this study. A predicted protein name for the putative product was assigned for each of the 6,975 CDS employing nBLAST from NCBI. Best matches were stored as plausible coded protein names.

5.2.2 Pleiotropy

The pleiotropy was approximated for every CDS of the emperor penguin by the search of Gene Ontology (GO) terms from the biological processes' ontology. The putative amino acid sequences of each CDS were obtained by translating the nucleotide sequences using Biopython's Seq module (Cock et al., 2009). To avoid missing GO terms associated with the target CDS because of the terms not being attributed directly to the emperor penguin's proteins, under the assumption of conservation of function, an orthologous search was performed to look for the GO terms associated with a whole orthologous group. This was performed using the Orthologous Matrix (OMA) database (Kaleb et al., 2019), which maps the protein sequence to hierarchical orthologous groups (HOG) in the database. Once the HOG membership was defined, the set of GO terms from the biological process ontology associated with each ortholog of the target sequence was retrieved. To avoid biological processes that do not correspond to processes that occur in the penguin system, we downloaded the annotations for all Spheniscidae (143,622 annotations) from QuickGO (Binns et al., 2009) to use as a filter pool for the terms associated with every CDS, only the terms present in the filter pool are considered valid terms. The hierarchical structure of the GO ontologies allows the ancestor terms to be defined corresponding to broader biological processes and descendant terms which define increasingly specific processes as the distance to the root of the ontology increases. All level 2 (terms whose link to the root is mediated by a single other term) ancestors or descendants for every associated term were determined using the Python module GOATOOLS (Klopfenstein et al., 2018). To determine the total number of processes of level 2 associated with each CDS an adjacency matrix was constructed for every term assigned to a CDS, where a pair of terms have an adjacency value of 1 if they share at least one level 2 ancestor or descendant in common and a 0 if they do not. With the

adjacency matrix, a procedure of Markov clustering (Van Dongen, 2008) was performed with inflation and expansion values of 2 for each CDS, yielding the final number of clusters taken as the approximation of pleiotropy. As a result of this procedure, we obtained a reduction to 15.7% of the original number of associated terms for each CDS on average.

5.2.3 Evolutionary rate and selection signature

The molecular evolutionary rate was approximated using PAML (Yang, 2007), which has different programs to calculate the substitution rates using maximum likelihood models. For the gene age analysis (see Subsection 5.2.4), we used the CODEML program which implements codon substitution models to obtain the average values of d_N and d_S (number of non-synonymous substitutions over the number of non-synonymous sites and the number of synonymous substitutions over the number of synonymous sites, respectively) over the whole sequence and all the branches of the provided penguin phylogeny. For the pairwise comparison between penguin lineages, we used the YN00 program which calculates the d_N and d_S between each pair of sequences. Every evolutionary rate index was transformed by a Box-Cox procedure. For a conventional search of genes under positive natural selection, the CODEML results were used. As a conservative test, we calculated the likelihood difference between two models, M1a and M2a, where the second model assumes that there exist some sites that have a d_N/d_S ratio greater than 1 (positive natural selection). Every CDS where we rejected the model without positive natural selection, by a criterion based on the critical χ^2 value (Jeffares et al., 2015), was selected for the Gene Ontology Enrichment Analysis (GOEA) performed with GOATOOLS.

5.2.4 Gene age

To assign a putative age to each CDS a phylostratigraphic analysis was performed (Domazet-Loso et al., 2007) using the "phylostratr" R package (Arendsee et al., 2019). To visualize

the general pattern of change while accounting for the uneven sample size of each phylostratum, random groups of consecutive strata of size from 1 to 5 were determined and merged, and the resulting slope of the linear regression was positioned at the gene age of the older phylostratum.

5.2.5 Environmental data

Mean ocean conditions for each of the penguin species were inferred from their global distributions. Ocean geospatial gridded information was taken from BIO Oracle repository (version 2.2) with "SDMpredictors" and "Leaflet" R software packages. The following Present Time Ocean Surface layers were downloaded: mean temperature, mean current velocity, mean ice cover, mean ice thickness, mean phosphate, mean nitrate, max salinity, mean silicate, mean iron, mean chlorophyll, and mean net primary productivity. Species occurrences were downloaded from GBIF (Global Biodiversity Information Facility). A set of 10,000 random spatial independent occurrences per species at a resolution of arcdegrees with 2 digits were bound to the geospatial information conditions with the R package "spocc" and "raster". Occurrences with missing values were omitted. Lastly, the mean value among all occurrences per species for each of the ocean parameters was then calculated. To visualize the differences in oceanic environmental variables between species we performed a principal component analysis (PCA) of two dimensions.

5.2.6 Significant biological processes

For every pairwise comparison of lineages of the same genus based on the pairwise d_N, a separate linear regression for every set of genes that shared a common biological process was performed. After a false discovery correction, biological processes with a linear regression p-value less than 0.05 were retrieved irrespective of the sign of the slope. For the environmental analysis, a linear regression was performed between the absolute

difference in every given environmental variable and the slope of the linear regression between pleiotropy and evolutionary rate for every term. The biological processes with a p-value less than 0.05 were retrieved for a GOEA. The clustering and visualization of GO terms were performed using the GO-Figure! (Reijnders and Waterhouse, 2021) using a similarity threshold of 0.4.

5.3 Results

5.3.1 Temporal analysis

We detected the three stages of pleiotropy-evolutionary rate relation change along the evolutionary history of the penguin lineage through phylostratigraphic analysis (Fig. 5.1). The non-synonymous evolutionary rate of the CDS originated at most during the radiation of Spheniscidae show a negative trend when plotted as a function of their pleiotropy, though the p-value of the linear regression is 0.133. This result suggests that the CDS originated between this point and the diversification of Aves around 167 Mya are transitioning between the first two stages. All data points between the transition time and 1000 Mya either do not show a significant relation between pleiotropy and evolutionary rate or have a positive relationship. Specifically, the points showing a p-value smaller than 0.05 and a positive slope are the phylostrata corresponding to the diversification of Archelosauria, Gnathostomata, Vertebrata, and Deuterostomia. Figure 5.1 also suggests that genes older than around 800 Mya already reached the third and last stage of equilibrium, where the two older phylostrata have p-values smaller than 0.05 with a regression with a negative slope. No point, except the older phylostratum corresponding to the last universal common ancestor, showed a significant correlation between synonymous evolutionary rate and pleiotropy.

The analysis of the pleiotropy-evolutionary rate relation as a function of the time

of divergence between pairs of penguin species is achieved by calculating the pairwise evolutionary rate (Fig. 5.2). This analysis showed a negative relation between pleiotropy and evolutionary rate for pairs that diverged less than 10 Mya (pairs of the same genus). Meanwhile, at the four points in time corresponding to the divergence of Aptenodytes from the rest, Pygoscelis from the rest, Eudyptes from the rest and, finally, Spheniscus from Eudyptula, we found significant positive correlations between pleiotropy and pairwise evolutionary rate. Between the pairs that show a significant negative correlation, we found all pairs of banded penguins, except for the sister species Humboldt and Galapagos penguins. We also found significance for the pair of great penguins and for four pairs of crested penguins (erect-crested/fiordland, erect-crested/macaroni, erect-crested/eastern rockhopper and macaroni/northern rockhopper).

By the use of PCA, we created two principal components that together explain more than 77% of the variance for the oceanic conditions between penguin species. We can see three major relationships between variables, the ice-related variables, the three oxides, and the relation between net primary productivity and maximum salinity. By grouping the penguin species by genus, we can see that there is almost no superposition between genera, where the brush-tailed penguins are the ones most separated from the rest, meanwhile, the little penguin finds itself inside the Spheniscus polygon.

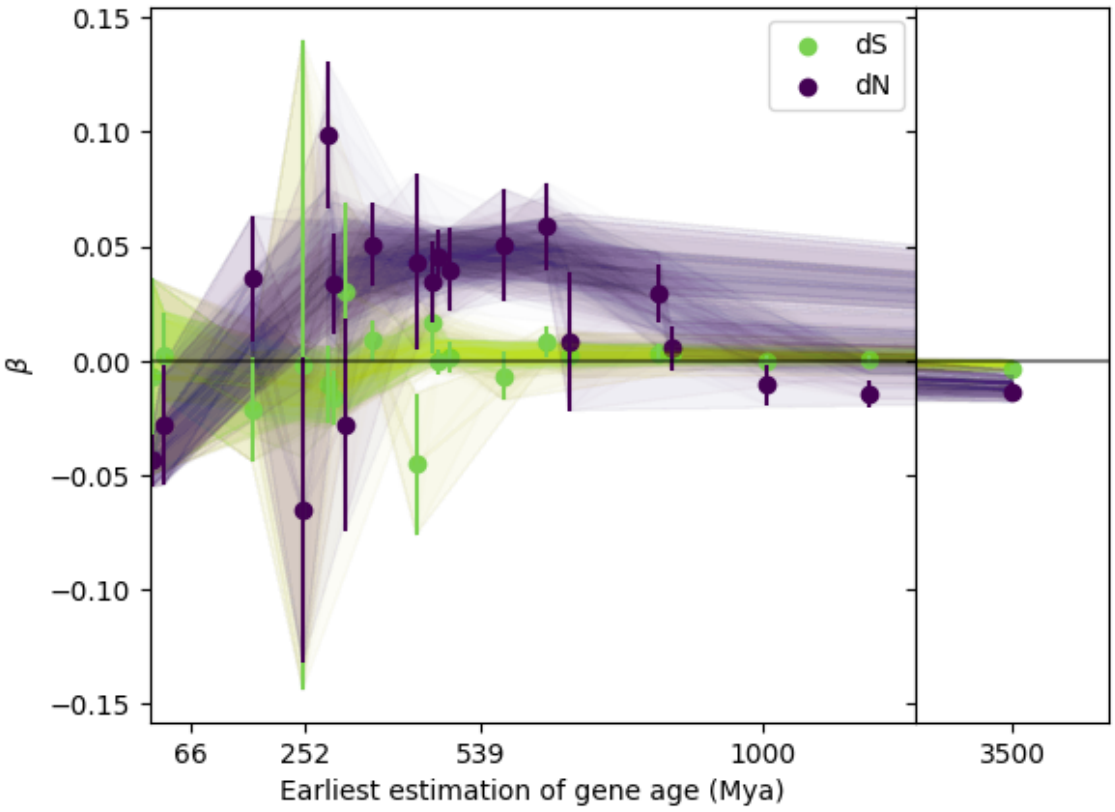

Figure 5.1: Pleiotropic effect on evolutionary rate as a function of gene age. Gene age is estimated as the divergence time of the oldest clade of each group of phylostrata, the time between the Eukaryota and Cellular Organisms phylostrata is skipped. The slope of the linear regression between the pleiotropy and the evolutionary rate (β) is calculated for synonymous (d_S) and non-synonymous (d_N) sites, corrected with a Box-Cox transformation. The error bars correspond to the standard deviation for the slope calculation of the linear regression. The results of the random group sorting are shown as translucent curves with filler between the error bars. 100 instances of group sorting are superimposed in the figure.

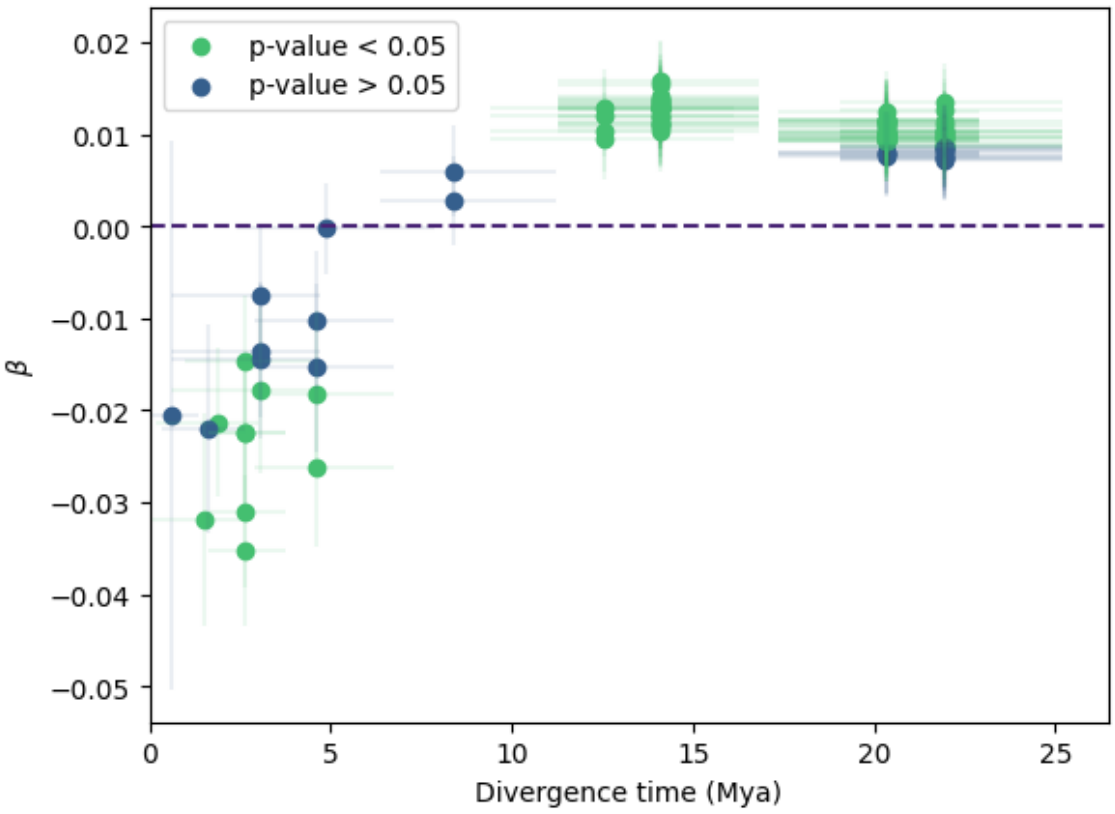

Figure 5.2: Pleiotropic effect on evolutionary rate as a function of divergence time between species. The slope of the linear regression between the pleiotropy and the evolutionary rate (β) is calculated for the non-synonymous sites corrected with a Box-Cox transformation. The vertical error bars correspond to the standard deviation for the slope calculation of the linear regression. The horizontal error bars correspond to the HPD 95% for the divergence time estimations. Different colors for the p-value of the linear regressions are used.

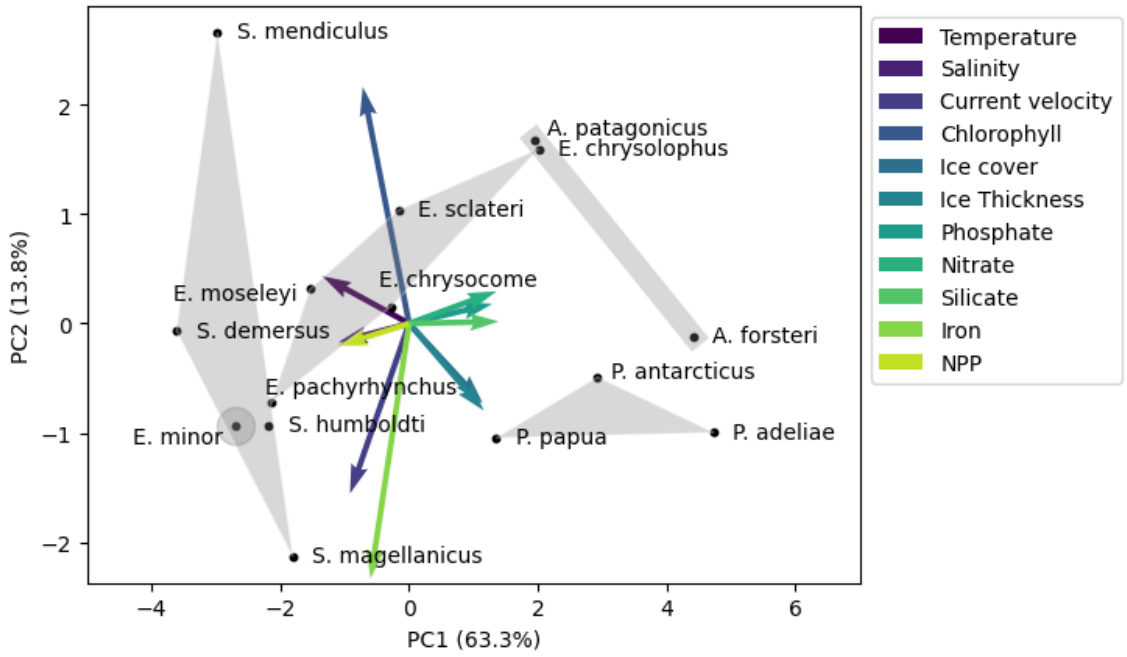

Figure 5.3: Bidimensional principal component analysis for 11 mean oceanic variables for the corresponding distributions of each of the 15 penguin species. The first and second component explains the 63.3% and 13.8% of the variance, respectively. The members of the same genus are shown with grey polygons.

5.3.2 Functional analysis

When comparing genes with signals of natural selection processes, we found that most of the biological processes enriched in CDS with high d_N/d_S are also found when looking for significant relationships between pleiotropy and evolutionary rate, this is the case for 8 groups of biological functions related to response to stimulus and gene expression, lipid metabolism, and hemostasis. On the other hand, only two groups of biological processes were found to be absent in the second analysis, which corresponds to the regulation of the immune response and sensory perception-related terms. Most of the biological processes shown were retrieved only from the pleiotropy-evolutionary rate analysis.

Using the linear regressions between pleiotropy and evolutionary rate in disaggregated data based on common biological processes, we determined the set of GO terms that show signs of evolutionary relevance in the pairwise intragenus comparisons. In the Spheniscus genus only five biological processes showed a greater sign of change, such as learning-related terms and glycosylation-related terms. For Aptenodytes, we found relevant changes in genes related to the development of the pancreas and chondrocytes, as well as in processes of glycosylation and response to starvation. We found for the Eudyptes genus an enrichment in relevant biological processes such as a variety of transport processes, cellular motor behavior, and ossification. In the case of the brush-tailed penguins, a big fraction of biological processes shows signs of greater change in the phenotypic optimum, where we find the processes related to heart contraction, catecholamine metabolism, and organic substance transport related to pigmentation.

An analysis of the relation between pleiotropy and evolutionary rate by environmental variable also suggests relevant biological processes. We found that for most of the environmental variables, about 10 to 20 groups of biological processes were formed using a similarity threshold of 0.4. Between the three points in penguin evolution analyzed, we found a lot of variability, where the most important variables changed drastically. For the differences found between the great penguins and the rest of the lineages, we found that the oxides, as well as the net primary productivity, are correlated with more groups of biological processes, meanwhile, the difference in chlorophyll levels hardly had any effect on any biological process. For the differences between crested penguins and banded and little penguins, we found much less biological processes, and a relative absence of effect of the oxides when compared to the effect of temperature and net primary productivity. This was also the only group that had an environmental variable with no effect on any biological process, maximum salinity. Finally, we found that the most prominent variables for the differences found between the brush-tailed penguins and the penguins

from Eudyptes, Spheniscus, and Eudyptula, are the ice-related variables (ice cover and thickness) which show almost 50 biological process groups affected significantly by the change in oceanic ice cover and thickness.

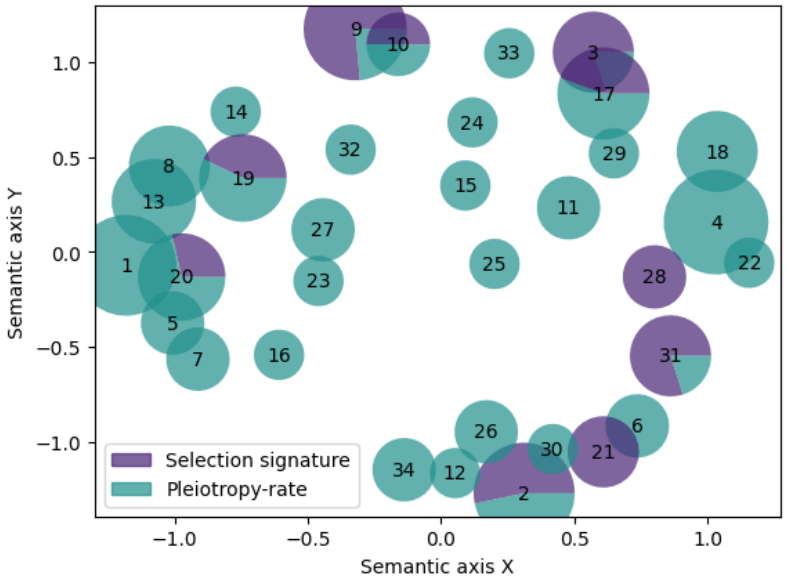

Figure 5.4: Comparison between the biological processes enriched in genes with high selection signal calculated conventionally by the d_N/d_S ratio and the biological processes that show a significant relationship between pleiotropy and evolutionary rate along the penguin evolutionary tree.

1. Protein biosynthetic, glycosylation and catabolic processes
2. Response and regulation of gene expression
3. Response to bacterium
4. Vesicle and protein localization and transport
5. Carbohydrate metabolic process
6. Regulation of catalytic activity
7. Nitrogen compound metabolic process
8. mRNA processing
9. Immunity, defense and response to external stimulus
10. Cellular response to amino acid stimulus
11. Cell cycle
12. Regulation of transcription elongation
13. Nucleic acid polymers repair and metabolism
14. Gene expression
15. Cellular process
16. Methylation
17. Chromosome and chromatin organization
18. Development
19. Cyclic nucleotide metabolic process
20. Lipid metabolic process
21. Regulation of immune response
22. Protein-containing complex localization
23. Electron transport chain
24. Cell division
25. Protein folding
26. Regulation of histone modification
27. Cellular catabolic process
28. Sensory perception
29. Chromosome segregation
30. Regulation of cell death
31. Hemostasis and coagulation
32. Cytoplasmic translational initiation
33. Membrane fission
34. Regulation of cell differentiation

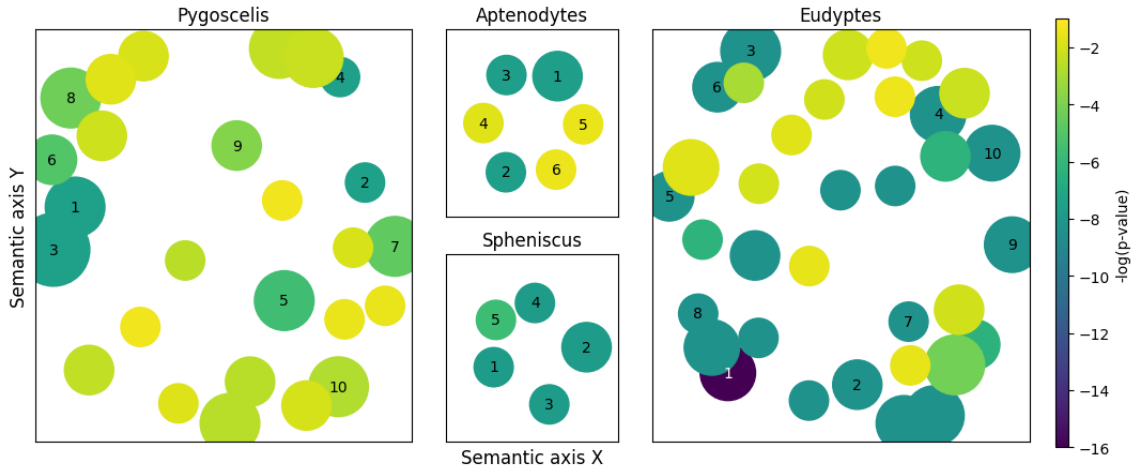

Figure 5.5: Biological processes that show a significant relation between pleiotropy and evolutionary rate for the pairwise comparison between penguins of the same genus. Each point corresponds to a group of biological processes with high semantic similarity. The color changes with the p-value of the linear regression, and the size corresponds to the number of biological processes grouped. The ten groups with the smallest p-value are labeled and shown below as their respective representative terms.

Pygoscelis
1. Positive regulation of heart contraction
2. Catecholamine metabolic process
3. Organic substance transport
4. DNA recombinase assembly
5. Amino acid biosynthetic process
6. Regulation of catalytic activity
7. RNA processing
8. Regulation of transcription by RNA polymerase II
9. Cell cycle
10. Protein ubiquitination

Aptenodytes
1. Chondrocyte development
2. Neutral amino acid transport
3. Fucose metabolic process
4. Response to starvation
5. Cellular process
6. Regulation of transcription by RNA polymerase II

Spheniscus
1. Regulation of neuronal synaptic plasticity
2. Ribosome disassembly
3. Hexose transmembrane transport
4. Regulation of neuron apoptotic process
5. Regulation of multicellular organismal process

Eudyptes
1. Transport
2. Regulation of mitochondrial membrane potential
3. Cilium organization
4. DNA replication
5. Bone morphogenesis
6. Endosomal vesicle fusion
7. Regulation of microtubule-based movement
8. Reverse cholesterol transport
9. Nuclear-transcribed mRNA catabolic process
10. Proteoglycan biosynthetic process

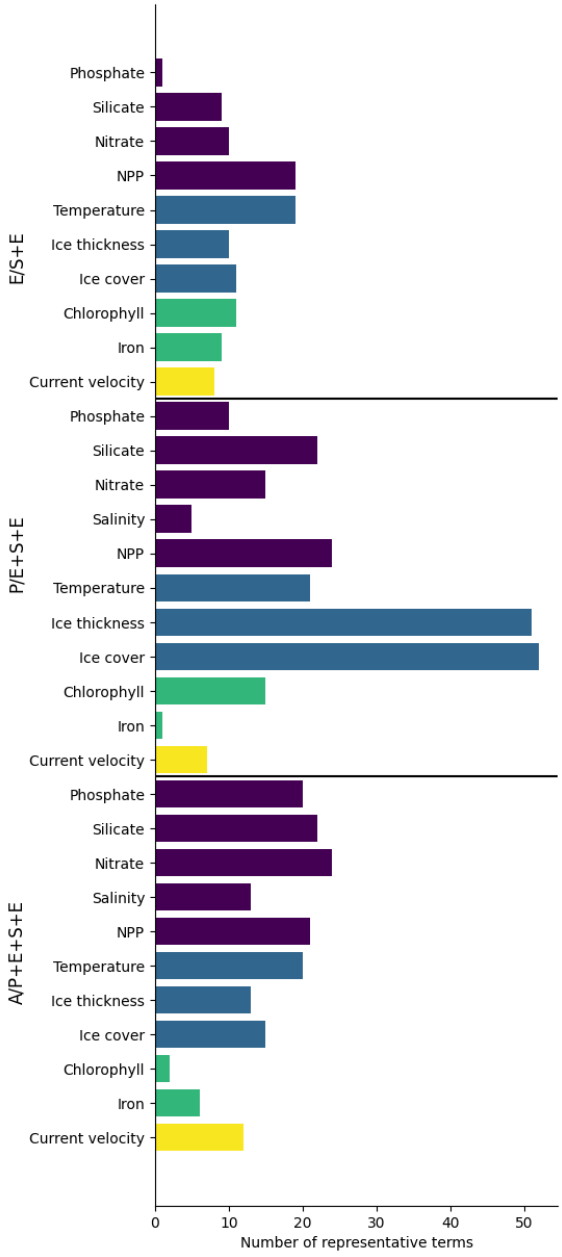

Figure 5.6: Number of biological processes that show a significant change of the pleiotropy-evolutionary rate relation as a function of environmental difference for 11 oceanic variables for comparisons of three groups. A/P+E+S+E compares pairs formed by one Aptenodytes penguin and any other penguin. P/E+S+E compares pairs formed by one Pygoscelis penguin with a penguin of Eudyptes, Spheniscus, or Eudyptula. E/S+E compares pairs formed by one crested penguin with one banded or little penguin.

5.4 Discussion

We examined significant milestones in penguin evolution aiming to investigate the changes in the pleiotropy-evolutionary rate relation. To ascertain the approximate phylogenetic origin of penguin coding sequences, we identified 20 major clades, referred to as phylostrata, all of them encompassing the emperor penguin as the focal lineage. The earliest phylostratum identified is the clade encompassing all three domains of life, where we did not differentiate genes that originated before the split of Bacteria from the genes that originated along the split between Archaea and Eukarya, finally encompassing genes that originated along a broad time window around 3.5 billion years ago. Each coding sequence was subjected to a comprehensive similarity search across the annotated genomes of 76 prokaryote organisms, the only phylostratum where more than 5 lineages were chosen. This phylostratum exhibited the largest number of associated CDS, with 4,704 protein sequences originating in this period, of which 3,520 were retained post-pleiotropy and evolutionary rate analyses. From that point on, the total number of CDS allocated in each phylostratum displayed a non-monotonic decline, dropping from 2,343 at the Eukaryota phylostratum, to the hundreds from the Opisthokonta to the Gnathostomata, to tens in the most recent phylostrata. Notably, an examination of the time intervals between phylostrata revealed that the periods encompassing the origins of animals and jawed vertebrates showed the highest number of coding sequence origin events per million years, with approximately 51 and 39 penguin genes originating per million years, respectively, results that lie under previous work on gene birth rates (Fernández and Gabaldón, 2020; Tan et al., 2021).

By linear regression analysis for each phylostratum, we managed to get an estimation of the effect of pleiotropy on the evolutionary rate for each CDS by age. The data suggest a confirmation of the expected behavior of the relationship, where we can see

the three stages with minor deviations (see Chapter 4). On one side, genes allocated at the diversification of crown Spheniscidae and genes present in the emperor but not in the king genome show a negative tendency (-0.028 and -0.044, respectively), with a p-value of 0.29 and 0.16, respectively, high p-values that can be attributed to the small sample sizes. On the other side, before a 1000 Mya, we do recover the second stage, with deviations from the positive tendency that can be attributed again to low sample sizes (Archosauria and Amniota with 5 and 15 CDS, respectively). Interestingly, the third stage is indeed significantly negative, but very close to zero, just as predicted by the model, where CDS with higher pleiotropy have a lower fitness equilibrium but experience a higher dimensional penalty on the evolutionary rate. As most effects of pleiotropy on evolutionary rate are mediated by the acceleration of mutation fixations by selective processes, it is to be expected that sites less prone to selection, as synonymous sites, have a weaker effect of pleiotropy on evolutionary rate, with a certain similarity to the relation encountered in non-synonymous sites at equilibrium where drift is the major contributor to the evolutionary process. As the effect is very close to zero relative to its variation, its statistical significance is more sensible to sample size, so is to be expected that we only find a small p-value on phylostrata with large sample sizes, as is the case of the phylostratum corresponding to cellular organisms.

If we interpret the results shown in Figure 4.9 as a reflection of the dynamics predicted by the FGM, as described in previous sections, we are to assume that, in general, the fastest evolving sequences take about 500 My to reach their optimum, meanwhile, the slowest take twice as much. Is hard to believe that a given sequence takes that long to reach its optimum, knowing that the processes of adaptation can be quite fast. To explain this, first is necessary to consider that the change in stage of relation between pleiotropy and evolutionary rate is given by the change from an adaptive walk state, dominated by selection processes, into the equilibrium state, dominated by the fixation of mutations

primarily by genetic drift. The extended adaptive walk observed can be caused by the continuous displacement of the optimum (Trubenová et al., 2019), instead of having a fixed phenotypic optimum as in the FGM simulations, this is most probably the case given that these sequences have been part of the genomes of multiple lineages, with very different ways of living. An alternative hypothesis is the shifting balance theory (Wright, 1932), which could explain the time dilation due to the exploration of multiple local maxima before reaching the global maximum. Although the fitness landscape changes along the multiple changes of the lineages, there is nonetheless a phenomenon of slow-down which appears to be affected by the pleiotropic level, in the same way as if the phenotypic optimum were fixed. Another important thing to note is that the evolutionary change recorded for each coding sequence corresponds to the change experienced during the last $\sim$21.9 My of penguin evolution, and the different phenotypic optimums related to their biological functions are recorded only for biological processes present in penguin systems, so any structural and functional change particular of a lineage different from penguins is going to be neglected by the analysis.

To see the first stage of negative correlation in more detail, we used the divergence dates of the species of penguin and their pairwise evolutionary rate. Effectively, there seems to be a clear effect of the evolutionary time on this relation as shown by Figure 5.2, where an initial negative relation is then succeeded by a positive relationship. The sister clades with the most recent divergence time correspond to the pair of banded penguins Humboldt and Galápagos, around 1.23 Mya, and even though the tendency is negative, not much of the genome has had time to change relative to other pairs, this means that we only have a few sequences to register the effect of pleiotropy on the evolutionary rate. This low sample size is most probably the reason for the high p-value of the linear regression. For the case of intergenic comparisons, we get positive correlations for every cladogenesis event, which is a sign that at least after around 12.58

Mya, the most pleiotropic sequences are already accumulating mutations faster than less pleiotropic ones. Interestingly, when comparing the means of the points obtained for the divergences of Eudyptula/Spheniscus and Eudyptes/Spheniscus+Eudyptula against Pygoscelis/Eudyptes+Spheniscus+Eudyptula and Aptenodytes/rest, we obtained that the first group statistically has a significantly greater mean than the second group. This could suggest that we are looking at the peak of the curve, which corresponds to the time of arrival of the less pleiotropic sequences to their equilibrium. This happens at around 15.42 My, which is significantly faster than the time seen in Figure 5.1 which is of the order of half a thousand million years. This discrepancy cannot be attributed to the difference between the mean non-synonymous substitution rate over site and branch and pairwise non-synonymous substitution rate, as the second is on average smaller than the former. Some hidden variables may exist that influence the evolutionary process between the penguins of the second group which buffers the effect of the pleiotropy over the evolutionary rate. One plausible hidden variable is the ecological distance between the penguins of the second group, relative to the ecological distance between the penguins of the first group. If the great penguins and the brush-tailed penguins have a greater ecological distance with Spheniscus, Eudyptula, and Eudyptes, than the difference within these last genera, then the sequences of the second group start their evolutionary trajectories further from the optimum, this shifts to the right the curve of stages, having the effect of appearing with a lower mean that an earlier group.

This suggestion is supported by Figure 5.3, which shows an approximate visualization of the differences between lineages as the differences in oceanic conditions where they are distributed. A greater ecological distance can be an approximate indicator of a greater change in the way of living of each lineage, which under the FGM corresponds to being further from the optimum, shifting the curve to the right. This is the case for Pygoscelis and Aptenodytes, which are secluded to the positive values of the first principal

component (involving mostly NPP, oxides, and temperature-related variables), far from Eudyptes, Spheniscus, and Eudyptula penguins, with the only exception being the strong resemblance between the environmental conditions of king and macaroni penguin. The difference between Eudyptes and Spheniscus is much less pronounced, meanwhile, Eudyptula is positioned inside the big Spheniscus polygon, indicating relatively mild differences in environmental conditions.

As a lineage changes its relationship with its environment, it is to be expected a shift in the particular combinations of trait values that yield the highest birth rates and lowest death rates. This optimum shift is greater as the change in the way of living of the lineage is greater, having effects on the accumulation of mutations in their different CDS. As in the FGM, this shift occurs along specific dimensions, corresponding to different phenotypic traits and biological processes, is to be expected that different sequences experience optimum shifts as long as the shift involves biological functions in which they participate. This is relevant because, from an equilibrium state of low fixation rate dominated by drift, a discrete set of CDS is going to be affected by an optimum change, starting a new evolutionary path away from the equilibrium. For a given time into the evolutionary process, the slope of the linear regression between pleiotropy and evolutionary rate is informative of the magnitude of the optimum shift, suggesting possible relevant biological functions that change the most during the evolution of a lineage. A huge change is going to throw the optimum very far from the phenotype of the population, starting the negative effect of pleiotropy on the evolutionary rate. Milder changes can move the optimum to distances where the first stage of the pleiotropy-evolutionary rate relation is negligible, so we visualize a positive effect of the pleiotropy of the sequences on their evolutionary rate. Finally, if the biological process stays the same and there is no phenotypic optimum shift, the sequences stay at the equilibrium zone, showing no noticeable relation between pleiotropy and evolutionary rate.

Using this reasoning, we retrieved all biological processes that show a significant effect of pleiotropy on the evolutionary rate when using the mean evolutionary rate across all sites and branches of the selected lineages. We found that no biological process showed a significant negative relation between pleiotropy and evolutionary rate, which suggests that this phase of the evolutionary process is relatively fast. To compare our analysis, we performed a conventional search of selection signature, detecting genes under positive selection and performing a GOEA to identify the relevant biological processes. We found a good superposition between both analyses, even though only one biological process term was shared between both (response to external stimulus), most of the processes detected by selection signature were grouped with processes with a significant effect of pleiotropy on the evolutionary rate. As did previous works (Vianna et al., 2020), we found selection signatures on functions related to blood, particularly related to response to injuries such as hemostasis and coagulation, but we did not find much evidence for this biological processes using the pleiotropy-rate method, where we only found processes broadly related to these, like calcium transport, the same occurs with immune and sensory-related functions, which are two general functions that are known to be consistently subjected to positive natural selection across the bird lineage (Shultz and Sackton, 2019). On the other side, in good agreement with previous studies, both analyses showed the relevance of processes involved in lipid metabolism, particularly important for this temperature-related clade and multicellular developmental processes, previously described to have strong selection signatures (Vianna et al., 2020; Pirri et al., 2022). For the most part, it seems that the pleiotropy-rate method is much less conservative, and highlights a broad array of cellular processes, a method that could be enriched highly by looking at the molecular differences between penguins and their sister group.

To extend this analysis further, we looked at the most relevant biological processes in the evolution of each of four genera, Aptenodytes (the great penguins), Pygoscelis

(brush-tailed penguins), Eudyptes (crested penguins) and Spheniscus (banded penguins). As shown in Figure 5.5 the results are highly influenced by the time of divergence of the lineages (in lineages that diverge further ago we are going to be able to detect more biological processes) and by the number of lineages (more lineages increase the number of comparisons and so the sample size, increasing the probability of significance). As a result, we found a lot more processes relevant in the evolution of Eudyptes and Pygoscelis, the former having a high number of species in the analysis and a long time of divergence and the latter having the oldest divergence times even though it has a small number of species. On the other side, Aptenodytes has only two species that diverged relatively recently, meanwhile, the four species of Spheniscus showed the most recent divergence times. Starting with Spheniscus, we found mainly two processes relevant to their evolution, the first one being processes related to the transport of sugars involved in glycosylation, fundamental for processes such as spermatogenesis, sperm maturation, competence between oocyte and sperm and fertilization, processes which are found to have changed drastically during the evolution of this group as shown in new studies (data not published). The second relevant processes that we found were related to learning processes such as the regulation of synaptic plasticity and rearrangement of nervous architecture. For the great penguins, we found relevant differences in pancreas development as well as response to starvation and sugar metabolic process, these results can shed light on the differences between the life cycle of king and emperor penguins, the latter living in much harsher conditions throughout the year. In Pygoscelis, we found many processes related to the metabolism of neurotransmitters like adrenaline, dopamine, and serotonin, as well as immunity-related processes similar to the ones detected by selection signal analysis. We also found that evolution under positive natural selection could have happened at genes related to the circulatory system, processes already determined to be subject to selection due to the diving behavior in cold waters of brush-tail penguins.

Finally, the crested penguins showed a huge number of relevant processes spanning a wide number of functions, like temperature homeostasis, mannosylation, and neurophysiological processes like forebrain development and sodium and potassium transmembrane transport, related to the propagation of the nerve impulse. Interestingly, we found many processes related to the tendency of rockhoppers to experience falls, such as angiogenesis and bone morphogenesis, as well as the balance-related process of inner ear development.

Finally, we investigated if some biological processes show a greater change in phenotypic optimum, suggested by a change in slope, as a function of a change in a given oceanic condition. To test this, we could not use all the pair comparisons at the same time because, as we saw earlier, the time of divergence between pairs affects greatly the change in slope. To avoid this effect, we made three comparisons each sharing a common divergence time, each comparison is characterized by the split of the penguin tree and the generation of an independent genus but is worth noting that the results are not particular to that given genus, but are results related to the difference between two specific groups of penguins. We don't see much difference in the variables of temperature, net primary productivity, and current velocity, all being fairly equal between the three groups and along the mean number of biological processes associated with them. For the case of the salts and oxides, we found a much greater relevance of these environmental variables for the A/P+E +S+E group and a much less significance for the E/S+E group. This last result is quite unexpected because Eudyptes penguins usually live in waters much richer in these compounds than Spheniscus and Eudyptula. The iron levels have consistently a lower number of biological processes associated which makes sense looking at Figure 5.3, as iron varies primarily along the second principal component, the axis along which there is not much phylogenetic disaggregation. Finally, we can see two linked major outliers in this analysis for the group P/E+S+E, which are the ocean ice variables, this result could be the effect of a strong phylogenetic signal, because of the shape of the curve

of these variables and the distribution of taxa along the curve, we can see that most of Pygoscelis lie at the far high levels of ice thickness and cover, with a fast descent into the levels experienced by other penguins, a lot of which are cero. This is not the case for most other variables, which show more evenly distributed values along the different clades and a milder slope, reducing the phylogenetic signal of the analysis.

Even though we found certain differences between environmental variables in biological process composition, those differences are not only due to the oceanic variable identity but they are also confounded by the phylogenetic conservation of traits. As an example, we found that the biological processes most directly related to the temperature at the A/P+E+S+E group that was effectively enriched for that variable and were not found for other variables were three terms related to fatty acid elongation, all other exclusive terms enriched for temperature in these group were terms not directly linked to temperature like defense response to virus, mature ribosome assembly, vesicle docking, reproductive structure development, and DNA repair. These results could be misleading as the number of lineages is small and the effects of ecology and evolution could be easily confounded.

6 Conclusiones generales

En esta tesis se estudió el efecto de la pleiotropía sobre la tasa de acumulación de mutaciones en simulaciones basadas en el modelo geométrico de Fisher y en secuencias codificantes de especies de la familia Spheniscidae. Se utilizó una amplia gama de herramientas computacionales incluyendo la resolución numérica de ecuaciones diferenciales, el modelamiento de procesos evolutivos, la obtención de información a partir de bases de datos y el análisis bioinformático de información genómica. Con esta aproximación se encontró que la pleiotropía tiene un efecto sobre la tasa evolutiva que es dinámico en el tiempo y está estructurado en tres etapas, una etapa de efecto negativo, donde las secuencias menos pleiotrópicas acumulan mutaciones más rápido, una etapa de efecto positivo, donde las secuencias más pleiotrópicas se mantienen en la fase de caminata adaptativa mientras que las menos pleiotrópicas entran a la fase de equilibrio y una etapa de efecto nulo, donde todas las secuencias se encuentran en el equilibrio. Adicionalmente, se estudió el rol que cumple la selección natural y la deriva genética en esta sucesión de eventos y se extendieron los resultados para hacer un análisis funcional de la historia evolutiva de los pingüinos.

Los resultados de esta investigación apoyan el rol determinante que tiene la pleiotropía sobre la tasa con la cual se incorporan cambios moleculares a los linajes. La participación de los productos proteicos en un elevado número de procesos biológicos va a restringir a corto plazo su tasa evolutiva por medio de la reducción en la porción de nuevas mutaciones de carácter beneficioso. La fase de caminata adaptativa está caracterizada por la contribución predominante de la selección natural direccional y purificadora, la fijación de un mayor porcentaje de mutaciones beneficiosas y un incremento sostenido del fitness de

la población (relativo a los rasgos particulares estudiados). En el caso de las secuencias codificantes de los pingüinos, los resultados bioinformáticos y de modelamiento sugieren que esta fase se extiende por una escala temporal de millones de años, lo cual puede ser indicio de óptimos fenotípicos móviles. A largo plazo, la disminución progresiva de la tasa evolutiva y la presencia de una zona de equilibrio dominada por la deriva genética, va a producir que las secuencias de menor pleiotropía comiencen a acumular mutaciones más lento que las secuencias más pleiotrópicas que siguen en la fase de caminata adaptativa. Finalmente, llega un momento en el proceso evolutivo de las secuencias donde ha pasado tanto tiempo que incluso las secuencias más pleiotrópicas han llegado al equilibrio, lo cual disminuye sustancialmente el efecto de la pleiotropía sobre la tasa evolutiva. Encontramos que un gran porcentaje del genoma de los pingüinos se encuentra en esta etapa. Adicionalmente, tomando en consideración los eventos recientes de especiación, los resultados apoyan un modelo donde, durante la especiación, ocurre el desplazamiento de los fenotipos óptimos para procesos biológicos específicos, desplazamiento que se puede inferir a partir de la etapa de la relación pleiotropía-tasa evolutiva en la que se encuentren secuencias codificantes que participen de tales procesos. Específicamente, pudimos rescatar para distintos géneros, por ejemplo, procesos vinculados con aprendizaje, vida en terrenos rocosos y conductas de riesgo, periodos de depleción de alimentos, reproducción y buceo en aguas heladas. Estos resultados apoyan la noción de que los supuestos asumidos por el modelo geométrico de Fisher son buenas aproximaciones a los procesos naturales o, alternativamente, que los resultados del modelo geométrico de Fisher son robustos a la modificación de sus supuestos básicos, como ya se ha descrito. Esto es interesante, ya que la modelación nos permite acceder a niveles de explicación científica que no suelen estar a disposición en el estudio evolutivo de linajes como los trabajados acá. En particular, pudimos acceder a un entendimiento en detalle de los procesos de acumulación de mutaciones y cómo las fases de las trayectorias evolutivas, caminata adaptativa

y equilibrio, y la influencia dinámica de la selección natural y de la deriva genética, dan origen a las tres etapas. Por medio del análisis de difusión encontramos que la selección tiene principalmente un rol de acelerador del proceso evolutivo, sin embargo, este efecto causa que las secuencias entren antes en las zonas de equilibrio dominadas por deriva. En este modelo, encontramos que el rol de la selección purificadora en la determinación de la tasa evolutiva es muy reducido, causando que encontremos que la tasa evolutiva mínima, experimentada por las secuencias en equilibrio, coincida en general con la tasa evolutiva neutral.

Este trabajo abordó un desafío central para la biología evolutiva en esta era desbordante de datos, el diálogo entre los estudios teóricos y la información de sistemas biológicos reales. Pudimos implementar satisfactoriamente un modelo de evolución molecular y utilizar las predicciones y explicaciones derivadas de él para hacer un aporte a la investigación de la historia evolutiva de un linaje tan interesante como son los pingüinos.

7 Financiamiento

Esta tesis se pudo llevar a cabo gracias a la Beca de Fortalecimiento del Doctorado (BFD o BFP) por el año 2019 otorgada por la Facultad de Ciencias de la Universidad de Chile que cubrió mi primer año de doctorado, gracias a la beca ANID de Doctorado Nacional 2020 (folio: 21201994) que cubrió los tres años siguientes y gracias a la Beca de Arancel de la Escuela de Postgrado de la Facultad de Ciencias que cubrió mi último año de doctorado.

A Appendix

A.1 Fisher's Geometric Model

Fisher's geometric model was introduced by Ronald Fisher in 1930 in the section *The nature of adaptation* in his book *The Genetic Theory of Natural Selection*. His model is an abstraction of the process of adaptation of populations and can be divided into four major parts: population and phenotypic space, mutations, fitness landscape, and the fixation process.

$\overrightarrow{a}$ is the vector with n elements that define a species' phenotype in a given time situated inside an n-dimensional space ($\mathbb{R}^n$),

$$\overrightarrow{a} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

with n being the number of phenotypic traits.

The species can change its phenotype in a process of mutation-fixation, where new mutations ($\overrightarrow{m}$) arise randomly with a given maximum magnitude (R), corresponding to the maximum phenotypic distance (phenotypic difference) between a wild-type phenotype and a mutant phenotype. To ensure that any mutation is equally possible, the direction of the mutation ($\overrightarrow{z}$) is defined by selecting the change in each trait from a normal distribution with mean 0 and standard deviation 1 ($N(0,1)$). Then the actual magnitude

(r) is calculated by multiplying the maximum magnitude and a random number p drawn from a uniform distribution in the unit interval $(U(0,1))$ raised to the power of $1/n$. Therefore, the phenotype of a mutant $\overrightarrow{a}_{MT}$ from a given wild-type phenotype $\overrightarrow{a}_{WT}$ is updated as follows:

$$\overrightarrow{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}, z_i \sim N(0,1)$$

$$r = Rp^{1/n}, p \sim U(0,1)$$

$$\overrightarrow{m} = r\frac{\overrightarrow{z}}{||\overrightarrow{z}||}$$

$$\overrightarrow{a}_{MT} = \overrightarrow{a}_{WT} + \overrightarrow{m}.$$

The fitness of every state in the population is calculated in reference of an optimal phenotype $(\overrightarrow{o})$ based on a Gaussian decay as a function of the Euclidean distance between $\overrightarrow{a}$ and $\overrightarrow{o}$. Based on the fitness of the wild-type (w^{WT}) and mutant form (w^{MT}), a selection coefficient (s) is calculated,

$$w_{\overrightarrow{a}} = e^{-\frac{||\overrightarrow{o}-\overrightarrow{a}||^2}{2}}$$

$$s = \frac{w^{MT}}{w^{WT}} - 1.$$

Given the selection coefficient of each mutation and the population size (N), the probability of fixation of the mutation is calculated as follows,

$$q = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}.$$

A.2 Geometric Mutation

To calculate the probability that a mutation with a given effect in fitness arises, the area of the n-dimensional cap form by the intersection between the two n-balls shown in Figure A.1 is calculated.

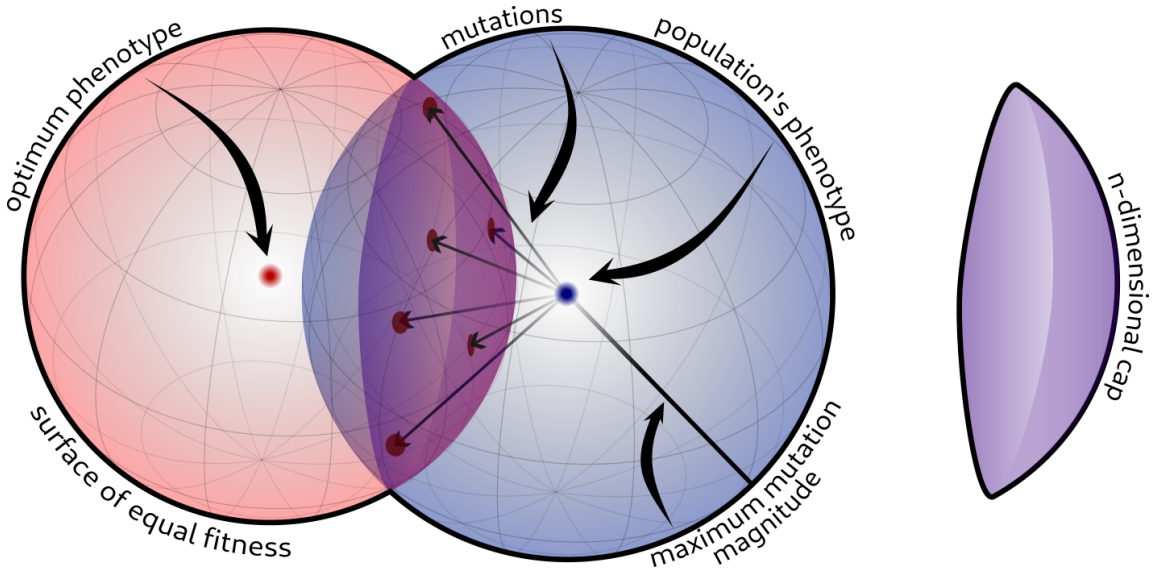

Figure A.1: Calculation of the mutation probability in Fishe's geometric model.

A.2.1 Cap height

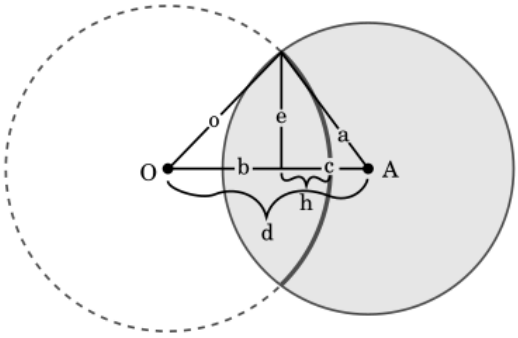

Figure A.2: Cap height for the smallest section. A two-dimensional slice of Figure A.1 is depicted. The segments a and o make a 90º angle. e is perpendicular to d, creating the segments b and c. h corresponds to the height of the cap and goes from the contact point between e and d to the intersection between d and the circumference centered at O.

The area's calculation requires knowing the height of the cap or the height of the complementary portion of the n-ball, depending on the distance between the species phenotype and the optimum phenotype, as well as the maximum mutation magnitude. The height of the cap (h) is calculated using the maximum mutation magnitude (a), the distance from the optimum of the mutant (o) and of the wild-type (d).

We have a triangle with sides a, o and d, where the height to d is e, which breaks d in two, b and c, so we have the following relation defining b and c

$$d = b + c. \tag{A.1}$$

Because e is the height on d, then, by the Pythagorean relation, we obtain

$$o^2 = b^2 + e^2 \tag{A.2}$$

$$a^2 = c^2 + e^2. \tag{A.3}$$

Merging equations A.2 and A.3 we get,

$$o^2 - a^2 = b^2 - c^2,$$

and so,

$$c = \sqrt{b^2 + a^2 - o^2},$$

which is used to reformulate the expression for b in equation A.1,

$$d = b + \sqrt{b^2 + a^2 - o^2}.$$

As $d - b$ is never negative we get

$$b = \frac{d^2 - a^2 + o^2}{2d}.$$

Finally, as h is equal to $o - b$ and, as b is the projection of o on d, $o \geq b$, then $h \geq 0$,

$$h = o - \frac{d^2 - a^2 + o^2}{2d},$$

then we get

$$h = \frac{a^2 - (d - o)^2}{2d}. \tag{A.4}$$

For the height of the complementary cap, the formula is derived as follows.

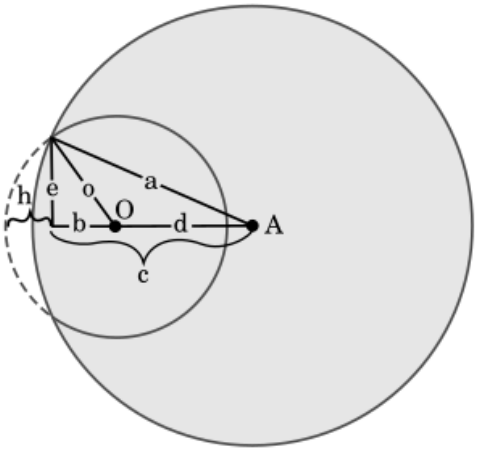

Figure A.3: Cap height for the biggest section. Two-dimensional representation of Figure A.1 if the center of the intersected n-ball lies within the intersecting n-ball. e is perpendicular to c, which is divided by the center of the intersected circle forming the segments b and d. h corresponds to the height of the cap denoted by the segmented line, defined by the intersection between both circles.

The height of the cap (h) is calculated using the maximum mutation magnitude (a), the distance from the optimum of the mutant (o) and of the wild-type (d), as follows,

$$c = d + b. \tag{A.5}$$

The Pythagorean relations hold,

$$o^2 = b^2 + e^2 \tag{A.6}$$

$$a^2 = c^2 + e^2. \tag{A.7}$$

Merging equations A.2 and A.3 we get,

$$o^2 - a^2 = b^2 - c^2$$

and so,

$$c = \sqrt{b^2 + a^2 - o^2},$$

which is used to reformulate the expression of c in equation A.5

$$d + b = \sqrt{b^2 + a^2 - o^2}$$

getting

$$b = \frac{a^2 - d^2 - o^2}{2d}.$$

Finally, as we can see that h is equal to $o - b$ and, as b is the projection of o, $o \geq b$, then $h \geq 0$,

$$h = o - \frac{a^2 - d^2 - o^2}{2d},$$

then we get

$$h = \frac{(d + o)^2 - a^2}{2d}. \tag{A.8}$$

A.2.2 Formulas for area and volume in hyperdimensional geometry

The volume of a n-ball of radius a is calculated as

$$V_{(a,n)} = \frac{\pi^{n/2}}{\Gamma\left[\frac{n}{2} + 1\right]} a^n,$$

where Γ is the gamma function. And the surface area of a n-ball of radius o is calculated as

$$A_{(o,n)} = \frac{2\pi^{n/2}}{\Gamma\left[\frac{n}{2}\right]} o^{n-1}.$$

The surface area of the cap of a n-ball of radius o with height h is

$$A^{cap}_{(o,n)} = \frac{1}{2}\frac{2\pi^{n/2}}{\Gamma\left[\frac{n}{2}\right]} o^{n-1} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right],$$

where I is the regularized incomplete beta function.

For the first case where the distance between the population and the optimum d is more than the maximum mutation magnitude a, the probability of mutation that changes the position of the population from d to o is

$$\begin{aligned} m_{d\to o} &= \frac{A^{cap}_{(o,n)}}{V_{(a,n)}} \\ &= \frac{1}{2}\frac{2\pi^{n/2}}{\Gamma\left[\frac{n}{2}\right]} o^{n-1} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right] \frac{\Gamma\left[\frac{n}{2}+1\right]}{\pi^{n/2}a^n} \\ &= \frac{\Gamma\left[\frac{n}{2}+1\right] o^{n-1}}{\Gamma\left[\frac{n}{2}\right] a^n} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right]. \end{aligned}$$

The gamma function has the following property

$$\frac{\Gamma[x+1]}{\Gamma[x]} = x,$$

so the final form is

$$m_{d\to o} = \frac{no^{n-1}}{2a^n} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right]. \tag{A.9}$$

If the d is less than a then we have three cases as shown in Figure A.4.

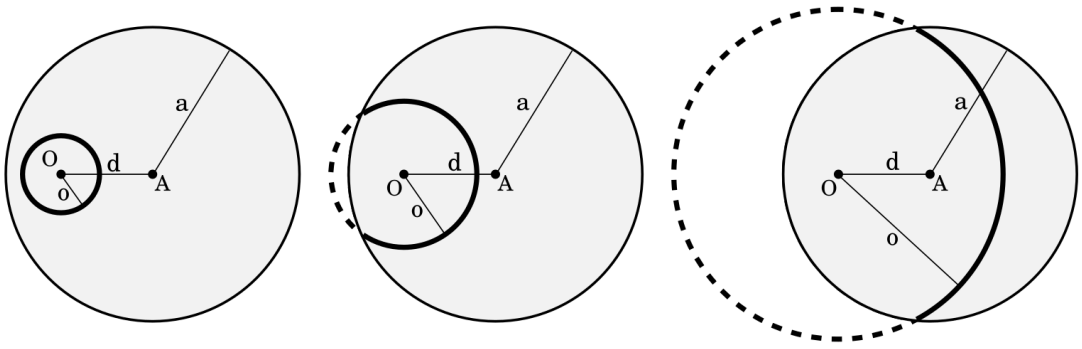

Figure A.4: Optimum distance cases

1. If $0 \leq o \leq |a - d|$, then

$$\begin{aligned} m_{d\to o} &= \frac{A_{(o,n)}}{V_{(a,n)}} \\ &= \frac{2\pi^{n/2}}{\Gamma\left[\frac{n}{2}\right]} o^{n-1} \frac{\Gamma\left[\frac{n}{2}+1\right]}{\pi^{n/2} a^n} \\ &= \frac{2\Gamma\left[\frac{n}{2}+1\right] o^{n-1}}{\Gamma\left[\frac{n}{2}\right] a^n}, \end{aligned}$$

and so,

$$m_{d\to o} = \frac{n o^{n-1}}{a^n}. \tag{A.10}$$

2. If $|a-d| \leq o \leq \sqrt{a^2-d^2}$, where $A^{cap}_{(o,n)}$ is the surface area of the small cap that is outside of the reach of the wild-type, so

$$\begin{aligned} m_{d\to o} &= \frac{A_{(o,n)} - A^{cap}_{(o,n)}}{V_{(a,n)}} \\ &= \left(\frac{2\pi^{n/2} o^{n-1}}{\Gamma\left[\frac{n}{2}\right]} - \frac{1}{2}\frac{2\pi^{n/2} o^{n-1}}{\Gamma\left[\frac{n}{2}\right]} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right] \right) \frac{\Gamma\left[\frac{n}{2}+1\right]}{\pi^{n/2} a^n} \\ &= \left(\frac{2\pi^{n/2} o^{n-1} - \pi^{n/2} o^{n-1} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right]}{\Gamma\left[\frac{n}{2}\right]} \right) \frac{\Gamma\left[\frac{n}{2}+1\right]}{\pi^{n/2} a^n} \\ &= \left(\frac{\pi^{n/2} o^{n-1} (2 - I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right])}{\Gamma\left[\frac{n}{2}\right]} \right) \frac{\Gamma\left[\frac{n}{2}+1\right]}{\pi^{n/2} a^n}, \end{aligned}$$

and so,

$$m_{d\to o} = \frac{n o^{n-1}}{a^n} \left(1 - \frac{1}{2} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right]\right). \tag{A.11}$$

3. If $\sqrt{a^2-d^2} \leq o \leq d+a$, the case is similar to the equation **??**, where we have

$$m_{d\to o} = \frac{A^{cap}_{(o,n)}}{V_{(a,n)}},$$

and so,

$$m_{d\to o} = \frac{n o^{n-1}}{2a^n} I\left[\frac{2oh-h^2}{o^2}; \frac{n-1}{2}, \frac{1}{2}\right]. \tag{A.12}$$

A.3 Fokker-Planck Equation

A.3.1 Diffusion analysis

The derivation of the Fokker-Planck equation proceeds as usual (see Gommes and Tharakan (2020) to see an example). Let's say that the position of a species is measured

as its Euclidean distance to the optimum phenotype, which behaves as a random variable $X(t)$, which depends on the occurrence and fixation of new mutations, which, in turn, depend only on the current position of the species, i.e. the process is memoryless in the Markovian sense. The probability that the species moves a distance ζ_i to the optimum depends on the probability that a mutation that causes such displacement occurs m_i (considering that multiple mutations can produce the same change in distance to the optimum) and on the probability that such mutation gets fixed in the population q_i (all mutations that make the same change in distance from the optimum have the same probability of fixation).

Then the random variable $X(t)$ obeys the following probabilistic law:

$$X(t+\tau) = X(t) + \partial X, \text{ where}$$
$$\partial X = \zeta_i \text{ with probability } m_i q_i.$$

This is true for every ζ_i except for $\zeta_i = 0$, which has a probability equal to $m_0 q_0$ plus the complement of the sum of the probability of every other possible displacement. That probabilistic law is only true if τ is the time interval equal to the expected time for a mutation to occur in an origin-fixation model. Consequently, the biggest step away from the optimum that a population can make in a τ interval corresponds to the maximum mutation magnitude (R). If the distance to the optimum is bigger than the maximum mutation magnitude, then the maximum value of ζ is R. If this is not the case, then the biggest step towards the optimum is the actual distance to the optimum, denoted here as d.

If $f(x,t)dx$ is the probability density function of the distance to the optimum, then:

$$f(x, t+\tau)dx = dx \int f(x-\zeta, t) m(x-\zeta, x) q(x-\zeta, x) d\zeta,$$

with the integral bounded between $-R$ and $min(d, R)$, which can be interpreted as if multiple populations were moving in Fisher's phenotypic space, then the number of populations in position x at time $t + \tau$ depends on the number of populations that are in position $x - \zeta$ at time t and that move a distance ζ in that given interval of time, considering all of those that stayed in position x.

Both the right-hand side and the left-hand side of the equation can be expressed approximately as a truncated Taylor series, specifically,

$$f(x, t + \tau) \simeq f(x, t) + \frac{\partial f}{\partial t}\tau,$$

$$f(x - \zeta, t)m(x - \zeta, x)q(x - \zeta, x) \simeq f(x, t)m(x, x + \zeta)q(x, x + \zeta) - \frac{\partial fmq}{\partial x}\zeta + \frac{1}{2}\frac{\partial^2 fmq}{\partial x^2}\zeta^2.$$

Given that

$$\int m(x, x + \zeta)q(x, x + \zeta) = 1,$$

the replacement on both sides yields the following equation:

$$\frac{\partial f}{\partial t} = \frac{1}{2\tau}\frac{\partial^2 f}{\partial x^2}\int \zeta^2 m(x, x+\zeta)q(x, x+\zeta)d\zeta - \frac{1}{\tau}\frac{\partial f}{\partial x}\int \zeta m(x, x+\zeta)q(x, x+\zeta)d\zeta. \quad \text{(A.13)}$$

A.3.2 Finite difference method

The equation is solved numerically by the finite difference method which requires the discretization of time and space. The space is divided into L equidistant points, each defined by the index i from $i = 0$ to $i = L - 1$. The terms A and B are defined so that equation A.13 can be rewritten as

$$\frac{\partial f}{\partial t} = A\frac{\partial^2 f}{\partial x^2} + B\frac{\partial f}{\partial x}.$$

The operators are approximated using the Crank-Nicolson scheme where i is the

spatial variable and j is the temporal variable:

$$\begin{aligned}
\frac{\partial f}{\partial t} &\approx \frac{f_{i,j+1} - f_{i,j}}{\Delta t}, \\
\frac{\partial^2 f}{\partial x^2} &\approx \frac{f_{i-1,j+1} - 2f_{i,j+1} + f_{i+1,j+1} + f_{i-1,j} - 2f_{i,j} + f_{i+1,j}}{2\Delta x^2}, \\
\frac{\partial f}{\partial x} &\approx \frac{f_{i+1,j+1} - f_{i-1,j+1} + f_{i+1,j} - f_{i-1,j}}{4\Delta x}.
\end{aligned}$$

Then the following factors are defined

$$\omega = \frac{1}{\Delta t}, \quad \alpha = \frac{A}{2\Delta x^2}, \quad \beta = \frac{B}{4\Delta x}.$$

The replacement results in the following formula

$$\begin{aligned}
\omega f_{i,j+1} - \omega f_{i,j} =& \alpha f_{i-1,j+1} - 2\alpha f_{i,j+1} + \alpha f_{i+1,j+1} \\
&+ \alpha f_{i-1,j} - 2\alpha f_{i,j} + \alpha f_{i+1,j} \\
&+ \beta f_{i+1,j+1} - \beta f_{i-1,j+1} + \beta f_{i+1,j}
\end{aligned}$$

and, by rearranging all $j+1$ terms to the left side and all j terms to the right side we get

$$\begin{aligned}
(-\alpha + \beta) f_{i-1,j+1} + (\omega + 2\alpha) f_{i,j+1} + (-\alpha - \beta) f_{i+1,j+1} & \\
= (\alpha - \beta) f_{i-1,j} + (\omega - 2\alpha) f_{i,j} + (\alpha + \beta) f_{i+1,j}.&
\end{aligned}$$

This is correct except for the borders of the system ($i = 0$, $i = L-1$), where $i = -1$ and $i = L$ lie outside of the defined discretized space. To get the correct expression for those two points a Neumann boundary condition (i.e. the derivative at the border is 0) is used. This is justified by the fact that close to the optimum we encounter an overshoot phenomenon and far from the optimum the probability of fixation of deleterious

mutations tends to 0. The derivatives are approximated by the forward and backward differences:

$$\begin{aligned}
\frac{f_0 - f_{-1}}{\Delta x} = 0 \rightarrow f_{-1,j+1} = f_{0,j+1}, f_{-1,j} = f_{0,j}, \\
\frac{f_1 - f_0}{\Delta x} = 0 \rightarrow f_{0,j+1} = f_{1,j+1}, f_{1,j} = f_{0,j}, \\
\frac{f_L - f_{L-1}}{\Delta x} = 0 \rightarrow f_{L,j+1} = f_{L-1,j+1}, f_{L,j} = f_{L-1,j}, \\
\frac{f_{L-1} - f_{L-2}}{\Delta x} = 0 \rightarrow f_{L-1,j+1} = f_{L-2,j+1}, f_{L-2,j} = f_{L-1,j}.
\end{aligned}$$

As a consequence, in $i = 0$:

$$\begin{aligned}
(-\alpha+\beta)f_{-1,j+1} + (\omega+2\alpha)f_{0,j+1} + (-\alpha-\beta)f_{1,j+1} = (\alpha-\beta)f_{-1,j} + (\omega-2\alpha)f_{0,j} + (\alpha+\beta)f_{1,j} \\
(-\alpha+\beta)f_{0,j+1} + (\omega+2\alpha)f_{0,j+1} + (-\alpha-\beta)f_{0,j+1} = (\alpha-\beta)f_{0,j} + (\omega-2\alpha)f_{0,j} + (\alpha+\beta)f_{0,j} \\
\omega f_{0,j+1} = \omega f_{0,j},
\end{aligned}$$

and for $i = L - 1$:

$$\begin{aligned}
(-\alpha+\beta)f_{L-2,j+1} + (\omega+2\alpha)f_{L-1,j+1} + (-\alpha-\beta)f_{L,j+1} = (\alpha-\beta)f_{L-2,j} + (\omega-2\alpha)f_{L-1,j} + (\alpha+\beta)f_{L,j} \\
(-\alpha+\beta)f_{L-1,j+1} + (\omega+2\alpha)f_{L-1,j+1} + (-\alpha-\beta)f_{L-1,j+1} = (\alpha-\beta)f_{L-1,j} + (\omega-2\alpha)f_{L-1,j} + (\alpha+\beta)f_{L-1,j} \\
\omega f_{L-1,j+1} = \omega f_{L-1,j}.
\end{aligned}$$

From these equations, we obtain a matrix expression on the function f for two points in time, where M_B and M_F correspond to the backward and forward matrices, respectively, defined as

$$M_B = \begin{bmatrix} \omega & 0 & 0 & \dots & 0 & 0 & 0 \\ -\alpha+\beta & \omega+2\alpha & -\alpha-\beta & \dots & 0 & 0 & 0 \\ 0 & -\alpha+\beta & \omega+2\alpha & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \omega+2\alpha & -\alpha-\beta & 0 \\ 0 & 0 & 0 & \dots & -\alpha+\beta & \omega+2\alpha & -\alpha-\beta \\ 0 & 0 & 0 & \dots & 0 & 0 & \omega \end{bmatrix},$$

and

$$M_F = \begin{bmatrix} \omega & 0 & 0 & \dots & 0 & 0 & 0 \\ \alpha-\beta & \omega-2\alpha & \alpha+\beta & \dots & 0 & 0 & 0 \\ 0 & \alpha-\beta & \omega-2\alpha & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \omega-2\alpha & \alpha+\beta & 0 \\ 0 & 0 & 0 & \dots & \alpha-\beta & \omega-2\alpha & \alpha+\beta \\ 0 & 0 & 0 & \dots & 0 & 0 & \omega, \end{bmatrix}$$

and so we have

$$M_B \begin{bmatrix} f_{0,j+1} \\ f_{1,j+1} \\ f_{2,j+1} \\ \vdots \\ f_{L-3,j+1} \\ f_{L-2,j+1} \\ f_{L-1,j+1} \end{bmatrix} = M_F \begin{bmatrix} f_{0,j} \\ f_{1,j} \\ f_{2,j} \\ \vdots \\ f_{L-3,j} \\ f_{L-2,j} \\ f_{L-1,j}. \end{bmatrix}.$$

With this expression and given an initial condition, for example, $f_{i_0,0} = 1$ for some i_0 between 0 and $L-1$, and $f_{i,0} = 0$ for every $i \neq i_0$, we can find the probability of finding the population for each spatial point along the whole spatial range in a given time t^*

$$\begin{bmatrix} f_{0,t^*} \\ f_{1,t^*} \\ f_{2,t^*} \\ \vdots \\ f_{L-3,t^*} \\ f_{L-2,t^*} \\ f_{L-1,t^*} \end{bmatrix} = (M_B^{-1} M_F)^{t^*} \begin{bmatrix} f_{0,0} \\ f_{1,0} \\ f_{2,0} \\ \vdots \\ f_{L-3,0} \\ f_{L-2,0} \\ f_{L-1,0} \end{bmatrix}.$$

Bibliography

Alvarez-Ponce, D. (2020). Richard Dickerson, molecular clocks, and rates of protein evolution. *Journal of Molecular Evolution*, 89:122–126.

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K., and Wurtele, E. (2019). phylostratr: a framework for phylostratigraphy. *Bioinformatics*, 35:3617–3627.

Arendt, W. and Urban, K. (2020). *Partial differential equations: an introduction to analytical and numerical methods*. Springer Spektrum.

Bailey, S., Alonso-Morales, L., and Kassen, R. (2021). Effects of synonymous mutations beyond codon bias: the evidence for adaptive synonymous subtitutions from microbial evolution experiments. *Genome Biology and Evolution*, 13:evab141.

Bataillon, T. and Bailey, S. (2014). Effects of new mutations on fitness: insights from models and data. *Annals of the New York Academy of Science*, 1320:76–92.

Beatty, J. (2016). The creativity of natural selection? Part I: Darwin, darwinism, and the mutationists. *Journal of the History of Biology*, 49:659–684.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, 25:3045–3046.

Blanquart, F., Achaz, G., Bataillon, T., and Tenaillon, O. (2014). Properties of selected mutations and genotypic landscapes under Fisher's geometric model. *Evolution*, 68:3537–3554.

Blanquart, F. and Bataillon, T. (2016). Epistasis and the structure of fitness landscapes: are experimental fitness landscapes compatible with Fisher's geometric model? *Genetics*, 203:847–862.

Bogumil, D. and Dagan, T. (2010). Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biology and Evolution*, 2:602–608.

Brunet, T. and Doolittle, W. (2018). The generality of constructive neutral evolution. *Biology and Phylosophy*, 33.

Burch, C. and Chao, L. (1999). Evolution by small steps and rugged landscapes in the RNA virus ϕ6. *Genetics*, 151:921–927.

Casillas, S. and Barbadilla, A. (2017). Molecular population genetics. *Genetics*, 205:1003–1035.

Chakraborty, S. and Ghosh, T. (2013). Evolutionary rate heterogeneity of core and attachment proteins in yeast protein complexes. *Genome Biology and Evolution*, 5:1366–1375.

Chakraborty, S., Panda, A., and Ghosh, T. (2016). Exploring the evolutionary rate differences between human disease and non-disease genes. *Genomics*, 108:18–24.

Charmouh, A., Bocedi, G., and Hartfield, M. (2023). Inferring the distribution of fitness effects and porportions of strongly deleterious mutations. *G3 Genes—Genomes—Genetics*, 13:jkad140.

Chesmore, K., Bartlett, J., Cheng, C., and Williams, S. (2016). Complex patterns of association between pleiotropy and transcription factor evolution. *Genome Biology and Evolution*, 8:3159–3170.

Chevin, L., Martin, G., and Lenormand, T. (2010). Fiher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. *Evolution*, 64:3213–3231.

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423.

Cole, T., Zhou, C., Fang, M., Pan, H., Ksepka, D., Fiddaman, S., Emerling, C., Thomas, D., Bi, X., Fang, Q., Ellegaard, M., Feng, S., Smith, A., Heath, T., Tennyson, A., García, P., Wood, J., Hadden, P., Grosser, S., Bost, C., Cherel, Y., Mattern, T., Hart, T., Sinding, M., Shepherd, L., Phillips, R., Quillfeldt, P., Masello, J., Bouzat, J., Ryan, P., Thompson, D., Ellenberg, U., Dann, P., Miller, G., Boersma, P., Zhao, R., Gilbert, M., Yang, H., Zhang, D., and Zhang, G. (2020). Genomic insights into the secondary aquatic transition of penguins. *Nature Communications*, 13(3912).

Crank, J. and Nicolson, P. (1996). A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Advances in Computational Mathematics*, 6:207–226.

Darwin, C. and Wallace, A. (1858). On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London, Zoology*, 3:45–62.

Dickerson, R. (1971). The structure of cytochrome c and the rates of molecular evolution. *Journal of Molecular Evolution*, 1:26–45.

Doebeli, M., Ispolatov, Y., and Simon, B. (2017). Towards a mechanistic foundation of evolutionary theory. *eLIFE*, 6:e23804.

Domazet-Loso, T., Brajkovic, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *TRENDS in Genetics*, 23:534–539.

Drummond, D., Bllom, J., Adami, C., Wilke, C., and Arnold, F. (2005). Why highly expressed proteins evolve slowly. *PNAS USA*, 102:14338–14343.

Ericson, E., Pylvänäinen, I., Fernandez-Ricaud, L., Nerman, O., Warringer, J., and Blomberg, A. (2006). Genetic pleiotropy in *Saccharomyces cerevisiae* quantified by high-resolution phenotypic profiling. *Molecular Genetics and Genomics*, 275:605–614.

Ewens, W. (2004). *Mathematical Population Genetics*. Springer, NY.

Fernández, R. and Gabaldón, T. (2020). Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution*, 4:524–533.

Fisher, R. A. (1930). *The Genetic Theory of Natural Selection*. The Clarendon Press.

Fraïsse, C., Sala, G., and Vicoso, B. (2019). Pleiotropy modulates the efficacy of selection in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 36:500–515.

Frugone, M., López, M., Segovia, N., Cole, T., Lowther, A., Pistorius, P., Dantas, G., Petry, M., Bonadonna, F., Trathan, P., Polanoski, A., Wienecke, B., Bi, K., Wang-Claypool, C., Waters, J., Bowie, R., Poulin, E., and Vianna, J. (2019). More than the eye can see: genomic insights into the drivers of genetic differentiation in Royal/Macaroni penguins across the Southern Ocean. *Genetics*, 151:921–927.

García-Dorado, A., Monedero, J., and López-Fanjul, C. (1998). The mutation rate and the distribution of mutational effects of viability and fitness in *Drosophila melanogaster*. *Genetica*, 102:255–265.

Gommes, C. and Tharakan, J. (2020). The Péclet number of a casino: diffusion and convection in a gambling context. *American Journal of Physics*, 88:439–447.

Gordo, I. and Campos, P. (2013). Evolution of clonal populations approaching fitness peak. *Biology Letters*, 9.

Gros, P., Le Nagard, H., and Tenaillon, O. (2009). The evolution of epistasis and its links with genetic robustness, complexity and drift in a phenotypic model of adaptation. *Genetics*, 182:227–293.

Gros, P. and Tenaillon, O. (2009). Selection for chaperone-like mediated genetic robustness at low mutation rate: impact of drift, epistasis and complexity. *Genetics*, 182:555–564.

Gu, X. (2007). Evolutionary framework for protein sequence evolution and gene pleiotropy. *Genetics*, 175:1813–1822.

Harman, R. and V., L. (2010). On decompositional algorithms for uniform sampling from n-spheres and n-balls. *Journal of Multivariate Analysis*, 101:2297–2304.

Hartl, D. and Taubes, C. (1996). Compensatory nearly neutral mutations: selection without adaptation. *Journal of Theoretical Biology*, 182:303–309.

Haygood, R. (2006). Mutation rate and the cost of complexity. *Molecular Biology and Evolution*, 23:957–963.

He, X. and Zhang, J. (2006). Toward a molecular understanding of pleiotropy. *Genetics*, 173:1885–1891.

Hietpas, R., Jensen, J., and Bolon, D. (2013). Shifting fitness landscapes in response to altered environments. *Evolution*, 67:3512–3522.

Hodgkin, J. (1998). Seven types of pleiotropy. *International Journal of Developmental Biology*, 42:501–505.

Hwang, S., Park, S., and Krug, J. (2017). Genotypic complexity of Fisher's geometric model. *Genetics*, 206:1049–1079.

Jeffares, D., Tomiczek, B., Sojo, V., and dos Reis, M. (2015). *Parasite Genomics Protocols*. Humana New York.

Johansen, A. (2010). *International Encyclopedia of Education*. Elsevier.

Joyce, P. and Abdo, Z. (2018). Determining the distribution of fitness effects using a generalized Beta-Burr distribution. *Theoretical Population Biology*, 122:88–96.

Kaleb, K., Vesztrocy, A., Altenhoff, A., and Dessimoz, C. (2019). Expanding the Orthologous Matrix (OMA) programmatic interfaces: REST API and the *OmaDB* packages for R and Python. *F1000Research*, 8.

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *PNAS USA*, 41:144–150.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, UK.

Kimura, M. and Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature*, 229:467–469.

Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *PNAS USA*, 71:2848–2852.

King, J. and Jukes, T. (1969). Non-darwinian evolution. *Science*, 164:788–798.

Kingman, J. (1977). On the properties of bilinear models for the balance between genetic mutation and selection. *Mathematical Proceedings of the Cambrisge Philosophical Society*, 81:443–453.

Kingman, J. (1978). A simple model for the balance between selection and mutation. *Journal of Applied Probability*, 15:1–12.

Klopfenstein, D., Zhang, L., Pederson, B., Ramírez, F., Vesztrocy, A., Naldi, A., Mungall, C., Yunes, J., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., and Tang, H. (2018). GOATOOLS: a Python library for Gene Ontology analyses. *Scientific Reports*, 8:10872.

Kopp, M. and Hermisson, J. (2009). The genetic basis of phenotypic adaptation I: fixation of beneficial mutations in the moving optimum model. *Genetics*, 182:233–249.

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4:66–70.

Lourenço, J., Galtier, N., and Glémin, S. (2011). Complexity, pleiotropy, and the fitness effect of mutations. *Evolution*, 65:1559–1571.

Manna, F., Martin, G., and Lenormand, T. (2011). Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics*, 189:923–937.

Manrubia, S. and Cuesta, J. (2015). Evolution on neutral networks accelerates the ticking rate of the molecular clock. *Journal of the Royal Society Interfase*, 12.

Margoliash, E. (1963). Primary structure and evolution of cytochrome c. *PNAS USA*, 50:672–679.

Martin, G., Elena, S., and Lenormand, T. (2007). Distributions of epistasis in microbes fit predictions from a fitness landscape model. *Nature Genetics*, 39:555–560.

Martin, G. and Lenormand, T. (2006a). The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution*, 60:2413–2427.

Martin, G. and Lenormand, T. (2006b). A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution*, 60:893–907.

Maturana, H. and Mpodozis, J. (1992). Origen de las especies por medio de la deriva natural. *Revista del Museo de Historia Natural de Chile*, 46:1–48.

McCandlish, D., Epstein, C., and Plotkin, J. (2014). The inevitability of unconditionally deleterious substitutions during adaptation. *Evolution*, 68:1351–1364.

McCandlish, D. and Stoltzfus, A. (2014). Modeling evolution using the probability of fixation: history and implications. *The Quarterly Review of Biology*, 89:225–252.

McGee, L., Sackman, A., Morrison, A., PPierce, J., Anisman, J., and Rokya, D. (2016). Synergistic pleiotropy overrides the costs of complexity in viral adaptation. *Genetics*, 202:285–295.

Morozov, A. (2013). Modelling biological evolution: recent progress, current challenges and future direction. *Interface Focus*, 3:20130054.

Moutinho, A., Eyre-Walker, A., and Dutheil, J. (2022). Strong evidence for the adaptive walk model of gene evolution in *Drosophila* and *Arabidopsis*. *Nature Reviews Genetics*, 8:921–931.

Muller, M. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2:19–20.

Orr, A. (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, 52:935–949.

Orr, A. and Coyne, J. (1992). The genetics of adaptation: a reassessment. *The American Naturalist*, 140:725–742.

Orr, H. A. (2000). Adaptation and the cost of complexity. *Evolution*, 54:13–20.

Pavlicev, M. and Cheverud, J. (2015). Constraints evolve: context dependency of gene effects allows evolution of pleiotropy. *Annual Review of Ecology, Evolution and Systematics*, 46:413–434.

Peck, J., Barreau, G., and Heath, S. (1997). Imperfect genes, fisherian mutation and the evolution of sex. *Genetics*, 145:1171–1199.

Pegueroles, C., Laurie, S., and Albà, M. (2013). Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular Biology and Evolution*, 30:1830–1842.

Perfeito, L., Sousa, A., Bataillon, T., and Gordo, I. (2014). Rates of fitness decline and rebound suggest pervasive epistasis. *Evolution*, 68:150–162.

Pertierra, L., Segovia, N., Noll, D., Marttinez, P., Pliscoff, P., Barbosa, A., Aragón, P., Rey, A., Pistorius, P., Trathan, P., Polanowski, A., Bonadonna, F., Le Bohec, C., Bi, K., Wang-Claypool, C., González-Acuña, D., Dantas, G., Bowie, R., Poulin, E., and Vianna, J. (2020). Cryptic speciation in gentoo penguins is driven by geographic isolation and regional marine conditions: unforseen vulnerabilities to global change. *Diversity and Distributions*, 26:958–975.

Pirri, F., Ometto, L., Fuselli, S., Fernandes, F., Ancona, L., Perta, N., Di Marino, D., Le Bohec, C., Zane, L., and Trucchi, E. (2022). Selection-driven adaptation to the extreme Antarctic environment in the Emperor penguin. *Heredity*, 129:317–326.

Podder, S., Mukhopadhyay, P., and Ghosh, T. (2009). Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. *Gene*, 439:11–16.

Poon, A. and Otto, S. (2000). Compensating for our load of mutations: freezing the meltdown of small populations. *Evolution*, 54:1467–1479.

Pritykin, Y., Ghersi, D., and Singh, M. (2015). Genome-wide detection and analysis of multifunctional genes. *PLoS Computational Biology*, 11:e1004467.

Pál, C., Papp, B., and Hurst, L. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158:927–931.

Ram, Y. and Hadany, L. (2015). The probability of improvement in Fisher's geometric model: a probabilistic approach. *Theoretical Population Biology*, 99:1–6.

Rapp, B. (2017). *Microfluidics: Modeling, Mechanics and Mathematics*. Elsevier Inc.

Razeto-Barry, P., Díaz, J., Cotoras, D., and Vásquez, R. (2011). Molecular evolution, mutation size and gene pleitropy: a geometric reexamination. *Genetics*, 187:877–885.

Razeto-Barry, P., Díaz, J., and Vásquez, R. (2012). The nearly neutral and selection theories of molecular evolution under the Fisher geometrical framework: substitution rate, population size, and complexity. *Genetics*, 191:523–534.

Razeto-Barry, P. and Frick, R. (2011). Probabilistic causation and the explanatory role of natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42:344–355.

Razeto-Barry, P. and Maldonado, K. (2011). Adaptive *cis*-regulatory changes may involve few mutations. *Evolution*, 65:3332–3335.

Razeto-Barry, P. and Vecchi, D. (2017). Mutational randomness as conditional independence and the experimental vindication of mutational Lamarckism. *Biological Reviews*, 92:673–683.

Reijnders, M. and Waterhouse, R. (2021). Summary visualizations of Gene Ontology terms with GO-Figure! *Frontiers in Bioinformatics*, 1:638255.

Rennison, D. and Peichel, C. (2022). Pleiotropy facilitates parallel adaptation in sticklebacks. *Molecular Ecology*, 31:1476–1486.

Rice, D., Good, B., and Desai, M. (2015). The evolutionarily stable distribution of fitness effects. *Genetics*, 200:321–329.

Rice, S. (1990). A geometric model for the evolution of development. *Journal of Theoretical Biology*, 143:319–342.

Salathé, M., Ackermann, M., and Bonhoeffer, S. (2006). The effect of multifunctionality on the rate of evolution in yeast. *Molecular Biology and Evolution*, 23:721–722.

Sella, G. (2009). An exact steady state solution of Fisher's geometric model and other models. *Theoretical Population Biology*, 75:30–34.

Shultz, A. and Sackton, T. (2019). Immune genes are hotspots of shared positive selection across birds and mammals. *eLife*, 8:e41815.

Stoltzfus, A. and Cable, K. (2014). Mendelian-Mutationism: the forgotten evolutionary synthesis. *Journal of the History of Biology*, 47:501–546.

Tan, M., Redmond, A., Dooley, H., Nozu, R., Sato, K., Kuraku, S., Koren, S., Phillippy, A., Dove, A., and Read, T. (2021). The whale shark genome reveals patterns of vertebrate gene family evolution. *eLife*, 10:e65394.

Tenaillon, O. (2014). The utility of Fisher's geometric model in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, 45:179–201.

Tenaillon, O., Silander, O., Uzan, J., and Chao, L. (2007). Quantifying organismal complexity using a population genetic approach. *PLoS ONE*, 2:e217.

Trubenová, B., Krejca, M., Lehre, P., and Kötzing, T. (2019). Surfing the seascape: adaptation in a changing environment. *Evolution*, 73:1356–1374.

Vale, P., Choisy, M., Froissart, R., Sanjuán, R., and Gandno, S. (2012). The distribution of mutational fitness effects of phage ϕX174 on diffeerent hosts. *Evolution*, 66:3495–3507.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *Siam Journal on Matrix Analysis and Applications*, 30:121–141.

Velenich, A. and Gore, J. (2013). The strength of genetic interactions scales weakly with mutational effects. *Genome Biology*, 14:R76.

Vianna, J., Fernandes, F., Frugone, M. J., Figueiró, H., Pertierra, L., Noll, D., Wang-Claypool, C., Lowther, A., Parker, P., Le Bohec, C., Bonadonna, F., Wienecke, B., Pistorius, P., Steinfurth, A., Burridge, C., Dantas, G., Poulin, E., Simison, W. B., Henderson, J., Eizirik, E., Nery, M., and Bowie, R. (2020). Genome-wide analyses reveal drivers of penguin diversification. *PNAS*, 117:22303–22310.

Wagner, G., Pavlicev, M., and Cheverud, J. (2007). The road to modularity. *Nature Reviews Genetics*, 8:921–931.

Wang, Z., Liao, B., and J., Z. (2010). Genomic patterns of pleiotropy and the evolution of complexity. *PNAS USA*, 107:18034–18039.

Wang, Z. and Zhang, J. (2009). Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genetics*, 5:e1000329.

Weinreich, D. and Knies, J. (2013). Fisher's geometric model of adaptation meets the functional synthesis: data on pairwise epistasis for fitness yields insights into the shape and size of phenotype space. *Evolution*, 67:2957–2972.

Welch, J. and Waxman, D. (2003). Modularity and the cost of complexity. *Evolution*, 57:1723–1734.

Wilke, C. and Adami, C. (2001). Interaction between directional epistasis and average mutational effects. *Proceedings Biological Sciences*, 268:1469–1474.

Williams, A., Ngo, T., Figueroa, V., and Tate, A. (2022). The effect of developmental pleiotropy on the evolution of insect immune genes. *Genome Biology and Evolution*, 15:1–16.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the VI International Congress of Genetics*, pages 356–366.

Wright, S. (1955). Classification of the factors of evolution. *Cold Spring Harbor Symposia on Quantitative Biology*, 20:16–24.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24:1586–1591.

Zhang, J. (2023). Patterns and evolutionary consequences of pleiotropy. *Annual Review of Ecology, Evolution, and Systematics*, 54:1–19.

Zhang, J. and Yang, J. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16:409–420.

Zuckerkandl, E. and Pauling, L. (1962). *Horizons in Biochemistry*. Academic Press, NY.