UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS FÍSICAS

La modularidad en la evolución de la recombinación

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS (FÍSICA)

PRESENTA:

Diego Andrés Hartasánchez Frenk

DIRECTOR DE TESIS: Dr. Christopher R. Stephens Stevens

MIEMBRO DE COMITÉ TUTORAL: Dr. Denis Boyer
MIEMBRO DE COMITÉ TUTORAL: Dr. Octavio Miramontes Vidal

pcf
posgrado en ciencias físicas
u n a m

MÉXICO, D.F. 2011

UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS FÍSICAS

La modularidad en la evolución de la recombinación

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS (FÍSICA)

PRESENTA:

Diego Andrés Hartasánchez Frenk

DIRECTOR DE TESIS: Dr. Christopher R. Stephens Stevens

MIEMBRO DE COMITÉ TUTORAL: Dr. Denis Boyer
MIEMBRO DE COMITÉ TUTORAL: Dr. Octavio Miramontes Vidal

pcf
posgrado en ciencias físicas
u n a m

MÉXICO, D.F. 2011

Índice general

A dos sabios de barba blanca:
Jorge Vázquez y Luis de Sebastián

“El joven Charles Darwin no sabía qué hacer con su vida.
El padre lo estimulaba:
- Serás una desgracia para ti y para tu familia.
A fines de 1831, se fue.
Regresó a Londres después de cinco años de navegaciones por el sur de América, las islas Galápagos y otros parajes.
Trajo tres tortugas gigantes, una de las cuales murió en el año 2007, en un zoológico de Australia.
Volvió cambiado. Hasta el padre se dio cuenta:
- ¡Tu cráneo tiene otra forma!
No sólo traía tortugas. También traía preguntas. Tenía la cabeza llena de preguntas.”

Eduardo Galeano

Resumen:

La recombinación genética es un mecanismo molecular de importancia fundamental en la evolución natural. Siendo así, es pertinente considerar que ha estado sujeta a la selección natural y que la distribución no homogénea de eventos de recombinación a lo largo del genoma es consecuencia de ello. Partiendo de la hipótesis de bloques constructores propuesta en el contexto de los algoritmos genéticos y haciendo uso de herramientas propias de esta disciplina, en esta tesis se busca evidencia de que la evolución de la recombinación está relacionada con la estructura modular del genoma. La hipótesis de esta tesis es que los puntos en donde ocurre la recombinación se han ido modificando a medida que se han ido conformando regiones en el genoma con efectos favorables. Mediante simulaciones computacionales se analizaron distintos modelos de recombinación bajo diferentes paisajes de adecuación. Los resultados muestran que la recombinación en puntos aleatorios favorece la conformación de módulos de alta adecuación en etapas tempranas de la evolución. A medida que los módulos se van conformando, la recombinación preferencial en puntos entre dichos módulos, favorece la combinación de módulos adecuados. Un modelo de recombinación en donde se lleva a cabo esta transición, permite un aumento más rápido en la adecuación promedio de la población en relación a otros modelos. Además, dicho modelo es consistente con observaciones experimentales que muestran que los eventos de recombinación en organismos eucariontes hoy en día ocurren de forma preferencial en zonas no codificantes del genoma.

Abstract:

Genetic recombination is a molecular mechanism of fundamental importance in natural evolution. It is thus reasonable to consider that as such it has been under the influence of natural selection, counting among its consequences the inhomogeneous distribution of recombination events along the genome. With the Building Block Hypothesis of Genetic Algorithms as a starting point and making use of the tools developed for this discipline, we search for evidence that recombination is linked to the modular structure of the genome. This thesis' hypothesis is that the use of crossover points has been modified throughout time as regions of the genome have acquired functional forms with positive effects. A few models of recombination were simulated under different fitness landscapes and were analyzed. Our results show that random crossover favors the formation of modules with high fitness in early stages of evolution. As these modules begin to form, preferential crossover between modules favors the combination of high fitness modules. A recombination model that allows for this transition increases the average fitness of the population at a faster rate than other models. Furthermore, this model is consistent with evidence that crossover points in eukaryotes are mostly located in non-coding regions of the genome.

Capítulo 1

Introducción

Los seres humanos, como habitantes de este planeta y como seres pensantes capaces de hacernos preguntas acerca de nuestra existencia, estamos inevitablemente destinados a preguntarnos cómo se originó la vida y cómo miles de millones de años de evolución permitieron la emergencia de todos los seres vivos que habitaron y habitan hoy el planeta Tierra.

Científicos, como C. Darwin, dedicaron su vida a observar la naturaleza e hicieron inferencias notables que con el paso de los años han permeado en la sociedad y que hoy son aceptadas por numerosos sectores de la población mundial. Un método científico, pensado como la comprobación lógica de hipótesis a partir de observaciones y mediciones, ha dado como resultado descubrimientos magníficos que poco a poco nos han dado más vías de acceso a estas preguntas.

Afortunadamente, más vías de acceso han significado más posibilidades para adentrarse al estudio de la evolución para científicos de áreas otrora completamente ajenas a ella y entre ellas, como la física y la lingüística, por ejemplo. Desde la física, la formalización del estudio de la evolución se ha dado junto con el auge en el estudio de sistemas complejos adaptativos. En física, entendemos la evolución de un sistema físico como la modificación de una o varias variables que lo caracterizan, a lo largo del tiempo.

Un sistema adaptativo, sea biológico, social o económico, es aquél que se comporta de forma diferente dependiendo de las características del entorno y las propias. Cambios en el medio, acarrearán cambios en el sistema y viceversa, dando como resultado la evolución del sistema, de forma indefinida o hasta alcanzar un equilibrio. Todos ellos, en principio, pueden tratarse como sistemas físicos y puede estudiarse su evolución.

En esta tesis se estudiará la evolución de un sistema adaptativo muy particular: la evolución misma. Específicamente, se estudiará la evolución de la recombinación genética en seres vivos partiendo de conceptos propios de la genética de poblaciones y de los algoritmos genéticos. La recombinación, en términos generales, se define como la combinación de material genético proveniente de dos individuos para dar lugar al material genético de un nuevo individuo. La hipótesis principal de esta tesis es que la evolución de la recombinación se dio a

la par de la evolución de la modularidad en el genoma. Se presentará evidencia teórica y mediante simulaciones computacionales que muestre esta relación.

1.1. Antecedentes

La evolución molecular, desde el origen de la vida en la Tierra hasta el día de hoy, es sin duda uno de los procesos más complejos que conocemos. Tan solo imaginar cómo se ha llegado desde una molécula autoreplicativa hasta un genoma que contiene la información para generar a un ser humano es una actividad mental muy emocionante. Se sabe bastante acerca de los mecanismos moleculares y procesos biológicos detrás de la evolución molecular y el "resultado" es tan maravilloso que no parecería mala idea imitar a la naturaleza si queremos que un proceso que está evolucionando tenga un exitoso porvenir. Tenemos cierto conocimiento de los mecanismos y procesos que se están dando hoy en día y que se han dado en los últimos 2 mil millones de años, desde el origen de los primeros organismos eucariontes. Sin embargo, poco sabemos acerca de la evolución del genoma desde el origen de la vida hasta entonces.

La secuenciación del genoma de múltiples seres vivos ha mostrado que el genoma tiene una estructura modular, que va desde el nivel de nucleótidos, hasta las redes genéticas, pasando por codones, exones, genes y operones. Por otro lado, la evidencia experimental muestra que la recombinación genética en seres superiores no es aleatoria. En todos los niveles de descripción existen zonas con mayor incidencia de recombinación que otras. Por ejemplo, existen zonas dentro de los cromosomas, llamados *hotspots*, que muestran una mayor incidencia de recombinación homóloga; por otro lado, es mucho más común que se lleve a cabo la recombinación en zonas intergénicas y no adentro de genes; e incluso es más común que la recombinación intragénica se dé en intrones que en exones.

Más allá del hecho de que la recombinación no se da de manera equitativa en todas las partes del genoma, parece que el uso diferencial de puntos de cruza está íntimamente relacionado con la estructura misma del genoma. En este trabajo, se busca evidencia de que la evolución de la estructura modular del genoma se dio a la par de la evolución de la recombinación.

Partamos de un escenario plausible relativamente cercano al origen de la vida. Consideremos protoorganismos, con un material genético pequeño y sencillo desde el punto de vista estructral, con capacidad autoreplicativa, pero con pocas funciones más. Pensamos que en un principio, la recombinación generalizada[1] probablemente era aleatoria. La elección de puntos de cruza entre cadenas de nucleótidos (no necesariamente del mismo tamaño[2]) debió de haber permitido

[1] Al referirnos a recombinación generalizada estamos tomando en cuenta a cualquier tipo de intercambio genético dentro de un organismo (o protoorganismo) o entre organismos, es decir, no sólo a la recombinación homóloga o al entrecruzamiento de cadenas de genes sino también a la duplicación, la deleción, la transposición, la inversión, etc.

[2] Por simplicidad, todas las simulaciones hechas en este trabajo son con poblaciones con cadenas de longitud constante para toda la población, y módulos de tamaño constante en cada cadena. Sin embargo, un escenario más realista sería con cadenas de longitud variable y tamaños de módulos variables dentro de las cadenas.

una mejor exploración del espacio de configuraciones nucleotídicas posibles.

En algún momento, la mutación y la recombinación aleatoria entre cadenas debió haber conformado secuencias funcionales de nucleótidos, probablemente de longitud corta, que aumentaran la adecuación del organismo. Una vez construido un módulo funcional (con alguna función catalizadora de reacciones, por ejemplo), sería completemente contraproducente destruir dicha secuencia nucleotídica. Siendo así, sería beneficioso evitar que la recombinación se llevara a cabo en algún punto dentro de la secuencia, pues al recombinarse con material genético sin función, el módulo funcional, como tal, dejaría de existir. En cambio, la recombinación afuera del módulo, permitiría que éste se dispersara más rápidamente en la población. Por lo tanto, con el paso del tiempo y la conformación de más módulos funcionales, se irían seleccionando también a aquellos organismos que tuvieran mecanismos moleculares que hicieran menos frecuente la recombinación intramodular y más frecuente la recombinación intermodular.

La evolución de la recombinación habría ocurrido como consecuencia y a la par de la evolución del genoma y de la construcción jerárquica de módulos funcionales. La recombinación aleatoria temprana se habría transformado en una recombinación estructurada, obteniéndo así los individuos portadores de las secuencias asociadas a una mayor adecuación.

1.1.1. La genética de poblaciones

La genética de poblaciones es una disciplina que estudia la constitución genética de una población y cómo ésta cambia como función del tiempo. En particular, estudia el efecto que tienen las principales fuerzas evolutivas (selección natural, deriva génica, mutación y flujo génico) en la distribución de frecuencias alélicas, tomando en cuenta, además, factores como la estructura y subdivisión poblacional. Entre otras cosas, la genética de poblaciones busca descifrar cómo se ha llevado a cabo la evolución molecular, y explicar fenómenos como la adecuación y la especiación.

La genética de poblaciones se desarrolló a principios del siglo XX a partir de la reconciliación entre la teoría evolutiva de C. Darwin y A. R. Wallace, el redescubrimiento de la genética mendeliana y los modelos biométricos o bioestadísticos que intentaban aplicar razonamiento y modelos estadísticos a la biología.

Los principales forjadores de la genética de poblaciones fueron R. A. Fisher, J. B. S. Haldane y S. Wright. A partir de 1918, Fisher escribió varios artículos que culminaron con la publicación de su libro *The Genetical Theory of Natural Selection* en 1930 [9]. Fisher mostró cómo la genética mendeliana era, contrario a lo que pensaban muchos genetistas en esa época, completamente consistente con la idea de evolución guiada por la selección natural.

En los años veinte, Haldane analizó matemáticamente un caso real de la selección natural: el cambio de coloración en la palomilla *Biston betularia*, debido a la industrialización de la zona que habitaban[3]. Los modelos de Haldane

[3]Aunque el caso de *Biston betularia* ha sido usado como ejemplo canónico de que la selec-

mostraron que la selección natural podía actuar a un ritmo más rápido que el propuesto por Fisher.

Wright se concentró en la evolución de poblaciones aisladas e introdujo en 1932 el concepto de *paisaje adaptativo*, un paisaje con picos y valles que representaba todos los posibles estados de una población y a través del cual, las poblaciones podían desplazarse a lo largo del tiempo, alejándose de picos adaptativos por fenómenos como la deriva génica y acércandose a otros picos por efecto de la selección natural.

Fisher y Wright diferían en torno a la importancia relativa de la selección natural y la deriva génica. Esta discusión, entre otras y gracias al trabajo de gente como T. Dobzhansky, E. Mayr, G. G. Simpson, G. Malécot y G. L. Stebbins, entre muchos otros, culminó con la conformación de una nueva teoría evolutiva, conocida como *síntesis moderna de la evolución* o *síntesis neodarwiniana*.

1.1.2. Los algoritmos genéticos

A partir de 1950 varios científicos de la computación estudiaron sistemas evolutivos utilizando operadores inspirados en la evolución natural como método de optimización aplicado a la solución de problemas de ingeniería. En la década de los sesenta, I. Rechenberg y H. P. Schwefel introdujeron las "estrategias evolutivas" (del original alemán *Evolutionsstrategie*) para optimizar parámetros reales para aparatos. L. J. Fogel, A. Owens y M. Walsh desarrollaron el "cómputo evolutivo" (*evolutionary programing* en inglés), una técnica para elegir las mejores soluciones a tareas específicas [24].

En los años sesenta, J. H. Holland inventó los algoritmos genéticos (del inglés *genetic algorithms*). A diferencia de las estrategias evolutivas y el cómputo evolutivo, el propósito original de Holland no era resolver problemas específicos sino estudiar el fenómeno de adaptación tal como ocurre en la naturaleza y desarrollar formas para importar mecanismos de la selección natural a sistemas computacionales [15].

Durante los últimos años, las fronteras entre las estrategias evolutivas, el cómputo evolutivo y los algoritmos genéticos se han desdibujado de tal forma que se utiliza comúnmente el término "algoritmos genéticos" para referirse a métodos heurísticos[4] de búsqueda que imitan los procesos de la evolución natural para optimizar procesos y para encontrar soluciones óptimas a problemas complejos y es ésta la definición que se adoptará en este trabajo.

1.2. La cuestión de la representación

Lo primero que debemos hacer si queremos modelar un proceso evolutivo es elegir a un representante, es decir, elegir qué o quién va a evolucionar. La

ción natural puede ocurrir en periodos cortos de tiempo, actualmente existe cierta controversia al respecto, pues algunos autores consideran que el cambio en frecuencias alélicas se puede deber a otros factores [10].

[4]La palabra "heurístico" proviene del griego *εύρίσκω* que significa "yo encuentro" o "yo descubro".

elección de un representante no es sencilla, pues si pensamos en la vida en nuestro planeta, la evolución sucede a muchos niveles al mismo tiempo: evolución de individuos en una población, evolución de especies en un ecosistema y evolución de ecosistemas en la biósfera, por ejemplo. Para nuestro propósito, la mejor elección es la evolución de individuos en una población pues es la que se puede estudiar de forma más sencilla.

Necesitamos entonces elegir una forma para representar a esta población donde entendemos "población" como un conjunto de organismos de una misma especie que viven en la misma zona geográfica.

En genética existen dos maneras fundamentales para describir y diferenciar a dos organismos de la misma población: su *genotipo* y su *fenotipo*. El genotipo es toda la información contenida en el material genético de un organismo. La expresión de esta información influida por el medio ambiente, determina el fenotipo: el conjunto de características observables (estructurales, bioquímicas, fisiológicas y conductuales) de un organismo. Suena lógico que en lugar de codificar el fenotipo (que puede ser muy complicado), simplemente se utilice al genotipo para describir a los organismos. Obviamente esto deja de lado al medio ambiente, pero a éste lo podemos incorporar en el proceso evolutivo.

Nuestra elección de representante es pues una población de individuos de una misma especie, donde cada individuo estará representado por su genotipo. Es pertinente hacer una breve descripción de cómo está codificado el material genético de los seres vivos para poder por lo menos saber qué tan alejado de la realidad estará el modelo específico que elijamos para describir el genoma.

1.2.1. El genoma

La información genética de todos los organismos vivos que se conocen está codificada en moléculas de ácidos nucléicos, como el DNA y el RNA. Los ácidos nucléicos están formados por cadenas lineales de nucleótidos: moléculas orgánicas formadas por la unión covalente de un monosacárido de cinco carbonos (pentosa), una base nitrogenada y un grupo fosfato. Existen dos tipos de bases nitrogenadas: las púricas y las pirimídicas. La adenina (A) y la guanina (G) son las bases púricas, mientras que la citosina (C), la timina (T) y el uracilo (U) son las bases pirimídicas. El DNA contiene A, G, C y T; el RNA, A, G, C y U. Pares de bases nitrogenadas forman enlaces (la G con la C, la A con la T y la A con el U) y este hecho (entre muchos otros, por supuesto) es esencial para permitir la duplicación y transcripción de la información genética. La secuencia lineal de nucleótidos que conforman el DNA o el RNA de un organismo es toda información genética (estrictamente hablando) del organismo.

Ahora bien, las secuencias nucleotídicas están agrupadas a su vez formando genes. Los genes son secuencias lineales de nucleótidos que contienen la información necesaria para la síntesis de un polipéptido (generalmente una proteína) con una función celular específica. El gen es considerado la unidad tanto de información genética como de herencia. Los genes están organizados en cromosomas y el lugar físico que ocupa un gen en el cromosoma se conoce como *locus*. Al conjunto de genes (y por lo tanto, de cromosomas) que contienen toda la

información genética de un individuo se llama *genoma*.

Existen además otros niveles de organización del genoma que es importante mencionar ya que revelan una estructura modular en el mismo. En la síntesis de proteínas, una secuencia de DNA se transcribe a RNA y éste a su vez, se traduce a una proteína. Durante la traducción a proteínas, una secuencia de nucleótidos se traduce en una secuencia de aminoácidos. Los nucleótidos se leen de tres en tres y a cada triada o *codón* corresponde un aminoácido. Al haber cuatro bases nitrogenadas, existen 64 codones posibles (4^3), pero tan solo existen 20 aminoácidos canónicos. Por lo tanto, varios codones diferentes pueden corresponder a un mismo aminoácido. El código genético [5] es, pues, redundante, y la degeneración ocurre sobre todo en la tercera posición. Esta degeneración implica que hay sitios "sinónimos" (la tercera posición de cada codón) y sitios "no sinónimos" (la primera y segunda posiciones) en el genoma.

De esta manera, un gen consta de una secuencia de nucleótidos; cada tres nucleótidos forman un codón; y una secuencia de codones codifican para una proteína (generalmente). En eucariontes, los genes también tienen otra estructura interna modular, pues están conformados por regiones no codificantes (intrones) y regiones codificantes (exones). Cerca de los genes, se encuentran también secuencias reguladoras y promotoras, que también se pueden considerar módulos.

Un *operón* se define como una unidad genética funcional formada por un grupo o complejo de genes capaces de ejercer una regulación de su propia expresión por medio de los sustratos con los que interaccionan las proteínas codificadas por sus genes.

Un operón está formado por genes estructurales y factores de control. Los genes estructurales codifican proteínas (generalmente enzimas) que participan en alguna vía metabólica y su expresión está regulada precisamente por los factores de control: un factor promotor y un operador; la presencia del operador activa o desactiva al promotor permitiendo o inhibiendo la expresión de los genes estructurales.

Un operón es un ejemplo de una *red de regulación genética*: un conjunto de genes que están relacionados entre sí e involucrados en alguna función celular o vía metabólica específica.

1.2.2. ¿Cómo podemos representar a los genes en una simulación computacional?

Tenemos secuencias de nucleótidos que forman genes. Grupos de genes forman redes de regulación genética y se encuentran distribuidos en cromosomas que conforman el genoma de un organismo. Al estudiar fenómenos biológicos, recurrir a simulaciones computacionales para recrear en una computadora lo que vemos en "el mundo real" resulta una herramienta enormemente útil. Nos permite llevar a cabo análisis muy específicos de manera rápida y con validez estadística; nos permite explorar una enorme cantidad de escenarios; nos per-

[5] Se le llama "código genético" al código que identifica a cada codón con su correspondiente aminoácido.

mite tener un control fino de nuestros parámetros; y demás posibilidades. Para ello requerimos representar más fielmente al sistema biológico en cuestión. Por otro lado, requerimos que la simulación sea eficiente, rápida y que nos pueda brindar información clara. Aquí nos topamos con una contradicción: por un lado requerimos que la simulación tenga el mayor número de parámetros biológicos posible para hacerla lo más apegada a la realidad y por otro lado, queremos que sea eficiente, rápida y eficaz en sus predicciones. Es preciso entonces, encontrar un justo medio.

En este sentido, es importante tener en cuenta cuántos de los niveles de organización mencionados en la sección 1.2 se quieren representar en la simulación y qué tan fielmente se van a representar cada uno de ellos.

1.2.3. Nuestra representación

Como se mencionó en 1.2, como modelo hemos elegido a una población de individuos de una misma especie (es decir, individuos que generan descendencia fértil si se cruzan entre ellos). Nuestra población será de tamaño constante e igual a N para todo tiempo y no habrá traslape de generaciones.

Cada individuo estará representado por una cadena de L bits (cada bit podrá tener un valor de 0 ó 1). Esta cadena representará de alguna manera al genoma del individuo. En lugar de cuatro posibles nucleótidos en cada posición tendremos dos posibles bits y evidentemente simularemos genomas de longitud corta (32 bits, por ejemplo) en relación a cualquier genoma real.

Estas cadenas estarán divididas a su vez en un número determinado de bloques de longitud B. Todos los bloques serán de igual longitud y no habrá separación alguna entre un bloque y otro, por lo tanto, L deberá ser siempre un múltiplo de B. Los bloques representarían a los genes en su versión más simplificada posible, es decir, secuencias lineales, contiguas, de igual tamaño.

La única función de los bloques o genes será determinar la *adecuación*[6] del individuo al que pertenecen. Para ello habremos de definir un *paisaje de adecuación* (concepto íntimamente ligado al *paisaje adaptativo* propuesto por Wright - ver sección 1.1.1) que relacione todas las posibles secuencias de bits con una adecuación determinada. Por razones que más adelante se harán evidentes, usaremos paisajes de adecuación modulares, en los cuales la adecuación de cada individuo será la suma de las adecuaciones de cada uno de los bloques que lo conforman.

El modelo evolutivo se describirá con detalle en los capítulos 4 y 5, pero constará de la aplicación sucesiva de dos operadores: selección y recombinación. La selección es la elección (con un componente aleatorio) de los individuos con mayor adecuación de la población para ser los padres de la generación siguiente. La recombinación es precisamente la conformación de individuos de la nueva generación a partir de los padres. Ésta se lleva a cabo eligiendo un punto de la cadena e intercambiando las cadenas de ambos padres a partir de ese punto.

[6] *Fitness*, en inglés. Entre más adecuado sea un individuo, mayor será su probabilidad de sobrevivir y reproducirse.

Las nuevas cadenas resultantes serán los individuos de la nueva generación. En general, los modelos evolutivos incorporan probabilidades de recombinación; en este trabajo, la probabilidad de recombinar será siempre uno. Generalmente, también, los modelos evolutivos (como se comentará en el capítulo 4) incorporan también otro operador, la mutación, que consiste en la sustitución de un bit por otro, pero dados los propósitos de este trabajo, la probabilidad de mutar será siempre cero[7].

Lo que permite la aplicación de los operadores de selección, recombinación y mutación es el "movimiento" de los individuos en el paisaje de adecuación (donde cada punto representa un genotipo, y a cada genotipo está asociada una adecuación). Como se explicará en la sección 5.1, cada operador modifica de manera distinta a los genotipos. Por ejemplo, tanto la recombinación como la mutación alteran los genotipos, pero en general, la recombinación genera alteraciones más bruscas (grandes saltos en el espacio de adecuación), mientras que la mutación genera alteraciones finas (movimientos cortos en el espacio de adecuación). La razón por la cual se ha decidido no implementar la mutación en las simulaciones es para poder distinguir más claramente los efectos de la recombinación. De hecho, para la descripción de los genotipos y de la forma en que los operadores los alteran resulta muy útil (sobre todo para estudiar la recombinación) implementar un *coarse graining*.

1.3. Granulado grueso y esquemas

El término "granulado grueso" proviene del inglés *coarse graining*, cuya traducción más apropida sería "haciendo un granulado grueso" y se trata de un concepto bastante gráfico. La diferencia entre un granulado fino y un granulado grueso es que en este último, los granos son más grandes. Podemos identificar, por ejemplo, el grado de resolución de un aparato de medición con el tamaño de los granos y decir que entre más finos los granos, mayor el nivel de resolución o, entre más gruesos los granos, menor el nivel de resolución. Si se hace, pues, un *coarse graining* durante el análisis de cierto comportamiento, se hará un análisis con un granulado grueso, sin fijarse en los detalles.

Existen muchas formas de hacer un granulado grueso a la hora de analizar cualquier problema. En nuestro caso, una forma de hacerlo es analizar el comportamiento de esquemas en lugar del comportamiento de cadenas (genotipos) específicas.

Un *esquema* (*schema* en inglés) es una plantilla que identifica a un subgrupo de cadenas que tienen ciertas similitudes. En particular, el valor de los bits en determinadas posiciones de la cadena es idéntico. Un esquema se especifica determinando el valor de cada bit de la cadena: 0, 1 o *, donde el asterisco indica que puede ser tanto 1 como 0.

[7]Podría parecer que sin mutación, los organismos perderían la capacidad de evolucionar, pero como se expondrá a lo largo de este trabajo, uno de los efectos de la recombinación es modificar los genotipos, lográndose así en última instancia, los mismos efectos que la mutación, aunque en forma, la mutación y la recombinación difieren fundamentalmente.

Por ejemplo, puede indentificarse a un grupo de cadenas de ocho bits con el esquema 1**1*01*, que incluye a todas las cadenas que tienen un 1 en la primera, cuarta y séptima posición y un 0 en la sexta posición. El valor de la segunda, tercera, quinta y octava posiciones es irrelevante. El *orden* del esquema es el número N_2 de posiciones fijas en el esquema (cuatro en nuestro ejemplo), donde $N_2 < N$. Por lo tanto, el número de cadenas distintas que conforman al subgrupo deteminado por un esquema será siempre 2^{N-N_2} (16 en este ejemplo). Toda esquema también tiene una *longitud característica* ($\delta(H)$) que corresponde a la distancia entre la primer y la última posición específica (siete en nuestro ejemplo). La longitud característica es importante pues entre mayor sea, mayor será la probabilidad de que se rompa cuando sufre una recombinación.

El propósito, podríamos llamar "general", de hacer un *coarse graining* es reducir el número de grados de libertad efectivos en el problema. Cualquier metodología que pueda lograr esto, es en principio útil, sin embargo, para lograrlo es importante elegir adecuadamente los esquemas que maximicen la información reduciendo los grados de libertad efectivos.

Podría, de hecho, parecer que cambiar de la descripción de genotipos a la de esquemas, aumentaría el número de grados de libertad, pues al haber 3 diferentes estados posibles para cada bit en un esquema, hay 3^N esquemas diferentes, mientras que sólo existen 2^N cadenas binarias diferentes. Sin embargo, cada esquema incluye a muchas cadenas binarias.

Se le puede adjudicar una adecuación específica a cada esquema igual al promedio de las adecuaciones de las cadenas que conforman al subgrupo que representa. Por lo tanto, al utilizar un esquema, el granulado grueso nos permite no fijarnos en los detalles específicos de las cadenas que nos son irrelevantes y hacer énfasis en aquellos que sí nos importan (las posiciones fijas, el orden y la longitud característica del esquema) sin perder medidas reales del sistema (las adecuaciones de las cadenas).

Capítulo 2

El cómo, cuándo y dónde de la recombinación

La recombinación es un proceso de importancia fundamental en la evolución, tanto de sistemas naturales como de sistemas artificiales. Sin embargo, y como se describirá de forma detallada en la sección 2.2, desde la biología y la genética de poblaciones, no se ha podido dar una explicación satisfactoria del porqué de la recombinación. No se sabe cómo surgió, ni cómo evolucionó y sobre todo, tampoco se conoce cuál es la ventaja evolutiva que explique su presencia.

Podríamos estar frente un caso de una solución atorada (*lock-in solution*, en inglés): una solución, que a pesar de no ser la mejor, es la que se mantiene debido a lo inmensamente complicado que resultaría cambiarla. En ese sentido, podríamos considerar que la recombinación fue una forma inicial que se encontró para combinar material genético y que ahora es prácticamente imposible cambiarla. Sin embargo, parece que la recombinación no es una solución atorada sino un proceso que se ha moldeado, sujeto a la selección natural, a lo largo de miles de millones de años de evolución de los mecanismos que la llevan a cabo. En este trabajo se pretende mostrar que es posible que la distribución no homogénea de la recombinación que observamos hoy en día en organismos eucariontes (ver sección 2.5) sea precisamente consecuencia de la evolución natural y que en términos evolutivos, esta forma de recombinación resulta favorable.

En cómputo evolutivo, en el contexto de los algoritmos genéticos, la hipótesis de bloques constructores (BBH, por *Building Block Hypothesis*) propuesta por Holland [15] dio una explicación lógica e intuitivamente muy poderosa para explicar la ventaja evolutiva de la recombinación. Básicamente, la BBH explica que la recombinación es una excelente y efectiva forma para combinar soluciones parciales buenas y cortas y así conformar soluciones completas, óptimas y más largas. En la sección 2.6 se analizarán los pormenores de esta hipótesis, pero en términos generales, investigaciones posteriores evidenciaron que, por un lado, era imprecisa y, por otro, sobresimplificaba y no respondía las preguntas en torno a la recombinación. Fundamentalmente, no se puede decir que la recombinación

sea “buena” o “mala” en general, dado que la respuesta a esta pregunta depende fuertemente del estado de la población[1] en el momento en que actúa y del paisaje de adecuación sobre el cual actúa.

2.1. ¿Por qué existe la recombinación?

Intentar contestar por qué existe la recombinación es sin duda fundamental pues integra dos cuestiones elementales: cuál es el origen evolutivo de la recombinación y, por otro lado, cuál es la ventaja intrínseca de la recombinación para explicar su presencia.

El origen de la recombinación podría estar asociado a su ventaja intrínseca, pero también podría deberse a eventos moleculares completamente ajenos a ella. En cualquier caso, conocer la ventaja intrínseca de la recombinación no dejaría de ser relevante, pues independientemente de su origen, habría que explicar por qué la recombinación (como un rasgo) no ha sido eliminada por selección natural.

Es de suponer que la recombinación se originó en cierto momento y con ciertas características, pero que éstas han evolucionado hasta adquirir las características que le conocemos (y que le desconocemos) hoy en día. Siendo así, la unión de las dos preguntas anteriores nos podría ayudar a contestar una pregunta más: ¿cómo ha evolucionado la recombinación?

Tradicionalmente, si un rasgo ha sido seleccionado por selección natural, pensamos en que para ello debió de haberle brindado a su portador una ventaja adaptativa. En otras palabras, suponemos que su portador debió de haber estado mejor adaptado a su medio ambiente que sus competidores y que por eso sobrevivió y transmitió dicho rasgo a sus descendientes.

Si consideramos que el rasgo en cuestión la recombinación, y más específicamente, la forma en que ésta se lleva a cabo, no es claro *a priori*, cómo podríamos definir en qué consiste su ventaja adaptativa. De hecho, la cuestión es delicada pues la selección natural (y así se nos ha enseñado) actúa sobre el fenotipo y no sobre el genotipo. Al no tener ningún efecto directo sobre el fenotipo, la recombinación no debería estar sujeta a la selección natural. Para resolver este conflicto, se dice que la recombinación le confiere a su portador, no una ventaja adaptativa como tal, sino una ventaja evolutiva, un rasgo que le confiere al individuo la capacidad para “evolucionar” más rápidamente que sus competidores[2].

[1] El “estado de la población” es el conjunto de genotipos representados en la población en un momento dado y sus frecuencias respectivas.

[2] Parece que hemos regresado, sin querer, al lamarckismo suponiendo que un individuo tiene la capacidad de “evolucionar”. En realidad, vamos a considerar que la recombinación se selecciona a nivel poblacional o de grupo. Es decir, que le confiere a una población la capacidad de evolucionar más rápidamente que sus competidores. Para cuantificar la velocidad de evolución de una población podemos medir su adecuación promedio como función del tiempo.

2.2. ¿Cuál es la ventaja intrínseca de la recombinación para explicar su presencia?

R. A. Fisher (1930) y H. J. Muller (1932) [9, 25] propusieron teorías muy parecidas entre sí para mostrar la ventaja intrínseca de la recombinación. Imaginemos que existen dos poblaciones finitas (una con recombinación y la otra sin recombinación) en donde ocurren, como suelen ocurrir en poblaciones naturales, mutaciones con alguna ventaja selectiva. Estas mutaciones ocurrirán en diferentes loci y tendrán efectos distintos. Esperamos que por selección natural, algunas de estas mutaciones favorables se fijen. Sin recombinación, la única posibilidad para que dos mutaciones favorables se fijen es que una de ellas ocurra en un descendiente de un individuo donde ya había ocurrido otra mutación. De lo contrario, es imposible que ambas mutaciones se fijen pues estarían compitiendo entre ellas permanentemente. Por otro lado, en una población con recombinación, la fijación de mutantes en loci diferentes ocurrirá de manera más o menos independiente. Siendo así, las mutaciones favorables podrán en un momento dado aparecer en un mismo genoma por recombinación. La población con recombinación tenderá a evolucionar más rápidamente pues en ella se perderán muchas menos mutaciones favorables que en la población sin recombinación. Si tuviéramos entonces que seleccionar la más adecuada entre las dos poblaciones, elegiríamos la población con recombinación. Únicamente en el caso en que la población fuera tan pequeña o la tasa de mutación tan baja que cada mutación se fijara antes de que otra mutación ocurriera en la población serían equivalentes ambas poblaciones [4].

Después de estos trabajos, poco se ahondó en este tema hasta finales de los años sesenta, excepto por un par de artículos de Muller. Precisamente en uno de ellos (1964), Muller propuso un mecanismo conocido como "la matraca de Muller" (*Muller's ratchet*) [26]. Muller consideró el caso de una población finita sin recombinación donde ocurren mutaciones deletéreas[3]. Conforme pasa el tiempo, mutaciones deletéreas se comienzan a acumular en la población y siendo ésta finita, los individuos sin mutaciones deletéreas comenzarán a eliminarse de la población por deriva génica. Tarde o temprano, todos los individuos de la población tendrán por lo menos una mutación deletérea y a partir de este momento comienzará a "tronar la matraca" (*"the ratchet will begin to click"*). Por deriva génica comenzarán a perderse todos los individuos con sólo una mutación deletérea, luego los individuos con dos mutaciones deletéreas y así sucesivamente. De esta manera, la matraca de Muller lleva a la acumulación sucesiva e ilimitada de alelos deletéreos en la población. Por lo contrario, si se permitiera la recombinación, siempre se podrían combinar regiones del genoma sin mutaciones deletéreas, evitando así el efecto matraca. En 1974, J. Felsenstein se refirió a la matraca de Muller como lo que bien podría ser el efecto cuantitativo más importante de la recombinación [8].

Este mecanismo de Muller supone que no ocurren mutaciones reversas. Las

[3] Una mutación deletérea es aquella que le confiere al individuo que la porta una disminución en su capacidad para sobrevivir y reproducirse.

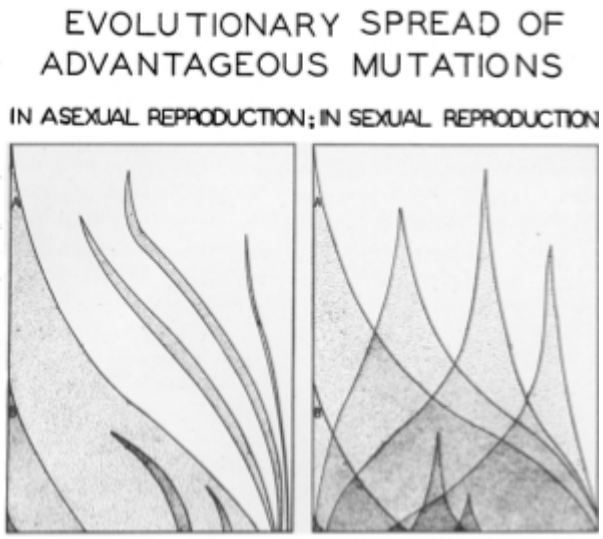

Figura 2.1: Tomada de [25]. Caricatura que muestra la propagación de mutaciones ventajosas en organismos asexuales (izquierda) y sexuales (derecha). El eje vertical corresponde al tiempo, que aumenta hacia abajo; el eje horizontal corresponde a una población dada, de tamaño constante. Las secciones de la población que portan los genes con mutaciones ventajosas son las obscurecidas de forma proporcional al número de tales genes que portan. En organismos asexuales estos genes compiten e impiden su mutua propagación; en organismos sexuales, se propagan a través de la población. [Pie de imagen del artículo original de Muller, con ligeras modificaciones. Traducción del autor de este trabajo.]

mutaciones reversas son aquellas que recuperan alelos previamente perdidos por una mutación deletérea y son altamente improbables pues sólo existe una manera (o muy pocas) para regresar un alelo a su forma original (i.e. la mutación contraria en el mismo lugar donde ocurrió la mutación deletérea) mientras que existen muchísimas maneras para que un alelo pierda su función.

En los años sesenta y setenta se suscitó cierta controversia alrededor de la existencia o no de una ventaja (intrínseca) asociada a la recombinación representada principalmente por J. F. Crow y M. Kimura por un lado y J. Maynard Smith por otro. Los detalles de esta controversia están muy bien plasmados por Felsenstein [8] en un artículo donde concluye que esta discrepancia se debió esencialmente a una diferencia en las suposiciones de los modelos. Mientras que los modelos que suponían poblaciones finitas mostraban una ventaja asociada a la recombinación, aquellos modelos que suponían poblaciones infinitas no encontraron tales ventajas si no existían epistasis ni desequilibrio de ligamiento inicial entre loci (estos conceptos se describirán en la sección 3.4).

Maynard Smith [21] hizo explícita la paradoja del doble costo del sexo (*the two-fold cost of sex paradox*), mostrando la desventaja asociada a la producción de gametos haploides por un organismo diploide[4].

[4]Un organismo diploide es aquél cuyas células somáticas contienen dos complementos cro-

Consideremos dos poblaciones: una asexual y la otra sexual. Por simplicidad, supongamos que cada hembra produce dos descendientes independientemente de si la reproducción es sexual o asexual. En el caso asexual, cada hembra producirá dos descendientes diploides (hembras) por partenogénesis, es decir, cada generación se verá duplicado el número de copias de su genoma. En el caso sexual, la hembra producirá un macho y una hembra haploides, cada uno con la mitad de su material genético. En este caso, la población se mantendrá constante, pues los machos sólo contribuirán a la siguiente generación fertilizando a las hembras.

Esto representa una desventaja del 50 % asociada a la producción de huevos haploides, es decir, que la población asexual crecerá el doble de rápido y la sexual acabará por desaparecer. Para contrarrestar el hecho de que "el sexo cuesta el doble", deben existir ventajas intrínsecas del sexo .

En 1966, W. G. Hill y A. Robertson usaron simulaciones computacionales para estudiar las interacciones entre varios loci en condiciones realistas, por ejemplo, bajo selección o con recombinación, en poblaciones finitas. Investigaron la probabilidad de fijación de una mutación benéfica en la presencia de otra mutación benéfica segregándose en un locus independiente. Confirmaron que en la presencia de deriva génica, la selección en un locus interfiere con la selección en un segundo locus independiente (sin epistasis), reduciendo su probabilidad de fijación. El "efecto Hill-Robertson" ocurre en poblaciones naturales de tamaño finito, y consiste en la reducción general en la efectividad de la selección debida al ligamiento entre sitios bajo selección [8]. Más recientemente, varios autores han mostrados que el efecto Hill-Robertson puede deberse a diversos fenomentos, entre ellos, *hitchhiking*, selección reversa y la acumulación de mutaciones deletéreas por la matraca de Muller [1].

Una ventaja intrínseca de la recombinación podría ser que ésta elimina el efecto Hill-Robertson. La interferencia entre loci genera desequilibrio de ligamiento negativo, es decir, la acumulación de alelos "buenos" en "contextos genéticos malos", que a su vez reduce la varianza genética en la adecuación, comparada con una población sin desquilibrio de ligamiento [2]. La recombinación incrementa la varianza en la adecuación de la población aumentando la respuesta a la selección [9]. Por lo tanto, si en una población ocurriera un mutación que incrementara la recombinación, se facilitaría la generación de descendientes con una mayor adecuación. La recombinción como tal tendría una ventaja selectiva indirecta y como tal, aumentaría su frecuencia (al estar asociada a genotipos más adecuados) [13].

2.3. Recombinación generalizada

Como se evidencia en la sección anterior, el problema del porqué de la recombinación aún no ha sido resuelto satisfactoriamente. Además, cabe resaltar que gran parte de la discusión histórica presentada aquí se ha centrado en la

mosómicos. Los organismos diploides pueden ser producto de la unión de dos gametos haploides, cada uno de los cuales únicamente contiene sólo un complemento cromosómico [10].

recombinación homóloga[5], sin embargo, la recombinación no se limita al entrecruzamiento de cromosomas homólogos. La *recombinación generalizada* incluye a muchas otras formas de recombinación, que incluyen la duplicación, la transposición, la deleción, la conversión génica, el entrecruzamiento desigual, etc.

La recombinación generalizada incluye a todos los mecanismos que utilizan al material genético existente para modificar una parte del genoma. En la sección anterior ya se comentó la universalidad de la recombinación (homóloga). La recombinación generalizada, en todas sus fascetas es por lo tanto, aún más ubicua, pues incluso se da en organismos como los virus, en los cuales no existen ni el sexo, ni los cromosomas.

A pesar de que en este trabajo únicamente simularemos y daremos la representación matemática de la recombinación homóloga (entrecruzamiento en un solo punto), cuando hablamos de la evolución de la recombinación, nos referimos a la evolución de la recombinación es su sentido más amplio, es decir, a la recombinación generalizada o el intercambio de material genético entre dos individuos, con la limitante de que las generaciones no se sobrelapan y que por lo tanto, no puede ocurrir intercambio de material genético entre individuos de diferentes generaciones.

En un principio, podría parecer que el proceso evolutivo que estamos caracterizando está sumamente simplificado, sin embargo, si tomamos en cuenta que la recombinación a la que nos referimos incluye a todo tipo de intercambio genético, el resultado es un proceso bastante general.

2.4. La recombinación homóloga y la *Holliday junction*

La recombinación homóloga, el intercambio genético entre cromosomas homólogos, se entendió (y se enseñó y sigue enseñándose a nivel básico superior) como la transferencia de información homóloga de un cromosoma a otro, esencialmente en la meiosis, aunque también existe la recombinación homóloga en la mitosis.

Con el avance de la biología molecular, la comprensión y el conocimiento de los mecanismos moleculares involucrados en la recombinación homóloga se han incrementado de forma importante. Además, el increíble avance en las técnicas de secuenciación de "próxima generación" (*NextGen*) y la correspondiente disminución en su precio, han permitido comprobar nuevas hipótesis y conocer más acerca de la evolución genómica.

En los últimos diez años se han propuestos múltiples modelos de la recombinación homóloga [16]. A pesar de que aún existe controversia respecto a cuáles son los modelos más cercanos a la realidad y que la incertidumbre respecto al porcentaje de eventos que se resuelven de una u otra forma es grande, la mayoría de autores coincide en que gran mayoría de eventos de recombinación

[5]De hecho, la discusión del porqué de la recombinación y la discusión del porqué del sexo, fueron una misma. Es claro que tienen cosas en común, pero es importante separarlas e históricamente no se hizo adecuadamente.

homóloga comienzan con un *double-strand break* (DSB)[6], es decir, un corte en cada una de las dos copias (duplicadas) de un cromosoma y se resuelven mediante la formación de una *Holliday junction*[7]. A continuación se describen los pasos que ocurren a partir de una DSB con dos resoluciones distintas: con y sin entrecruzamiento (ver figura 2.2).

1. *DSB*: ocurre un DSB en ambas copias de uno de los cromosomas duplicados del heteroduplex.

Llamaremos a estos cromosomas 1.1 y 1.2, siendo el 1.2 el más cercano al cromosoma homólogo.

2. *Resection*: Exonucleasas "se comen" las colas 5' de ambos cromosomas. Estas secciones no se regeneran pues las pues las colas 5' no crecen "naturalmente".

3. *Strand invasion*: ocurre una invasión del extremo 3' del cromosoma 1.2 al duplex homólogo que hasta ahora permanece igual. Este cromosoma (2.1) invade a su vez al primero formando un *D-loop*.

4. *Gene conversion*: se utiliza al cromosoma 2.2 como plantilla para reconstruir al cromosoma 1.2. Por otro lado, el cromosoma 2.1 sirve como plantilla para la reconstrucción del cromosoma 1.1.

5. *Resolution*: cada uno de los dos entrecruzmientos se puede resolver de dos maneras distintas:

a) que ocurran dos cortes ("horizontales"), de tal manera que se intercambien los fragmentos del medio de las cadenas entre los cromosomas 1.2 y 2.1, ó b) que de un lado ocurra un corte horizontal igual que en el caso a), pero que en el otro extremo ocurra un corte en los cromosomas 1.1 y 2.2, a la misma "altura" que el entrecruzamiento entre las cadenas 1.2 y 2.1 de tal forma que los cromosomas 1.1 y 2.2 intercambien una de sus colas.

A la resolución b) se le conoce precisamente como "entrecruzamiento". En el caso a) no ocurre un entrecruzamiento como tal, sino solamente conversión génica.

Este modelo propuesto originalmente como *Double-Strand Break Repair* (DSBR) por Szostak y colaboradores [38] consideraba que el número de resoluciones tipo a) debería ser igual a las de tipo b). Observaciones han sugerido que sólo el 10 % de las DSBs se resuelven mediante entrecruzamiento, con lo cual se han propuesto otros modelos como la *Synthesis-Dependent Strand Annealing* (SDSA) para resolver la problemática. En [28] se propone el 90 % de los DSBs se resuelven mediante SDSA sin entrecruzamiento y que el 10 % restante se resuelve mediante DSBRs sin entrecruzamiento, reduciendo a un mínimo la resolución mediante DSBRs y entrecruzamiento. En la SDSA, el *D-loop* descrito en el paso 3 se retracta y la cadena 1.2, junto con su nuevo segmento formado por conversión génica con la cadena 2.2, se une de nuevo con su cola, sin que ocurra un entrecruzamiento (ver figura 2.2).

A pesar de que tanto la DSBR como la SDSA son mecanismos que explican muchas observaciones experimentales, aún queda mucho por conocerse de los

[6] La mayoría de los términos utilizados en esta descripción son los utilizados comúnmente por la comunidad científica internacional en inglés.

[7] Algunos autores prefieren el término *Double Holliday junction*.

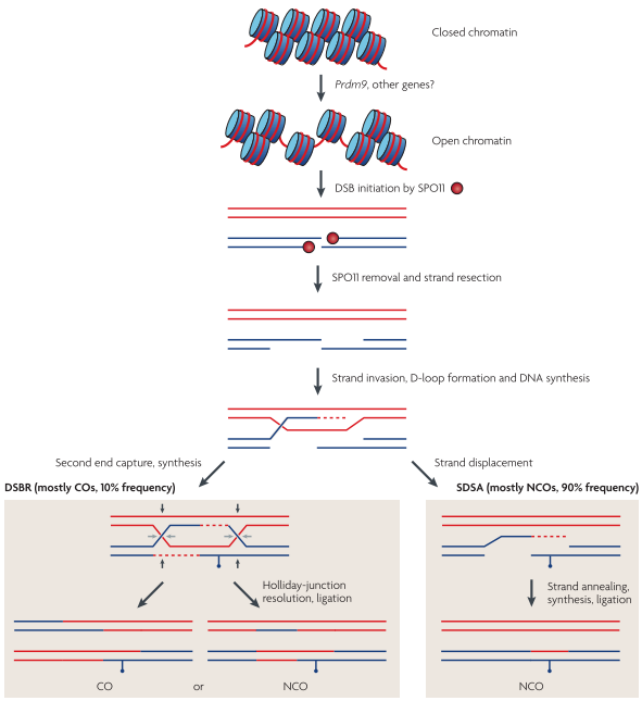

Figura 2.2: Tomada de [28]. La recombinación comienza con la activación local de la cromatina (por genes como el *Prdm9*), permitiendo que topoisomerasas (SPO11) catalicen un *Double-strand break* (DSB) en una de las cuatro cromátidas. La resección de las colas 5' deja a las colas 3' sueltas, que invaden a la cromátida no hermana, conformándose una *Holliday junction*. Los DSBs se reparan por un proceso de recombinación homóloga, mediante entrecruzamiento (CO), con el intercambio de marcadores laterales, o conversion génica sin entrecruzamiento (NCO) donde la cromátida donde se inició el DSB adquiere una secuencia corta de su pareja homóloga sin el intercambio de marcadores laterales. Los sitios donde se llevó a cabo el DSB se reparan utilizando la otra cromátida como templete. Se cree que los productos CO y NCO son resultado de dos vías distintas: *Double-Strand Break Repair* (DSBR), que da como resultado COs sobre todo y *Synthesis-Dependent Strand Annealing* (SDSA) que da como resultado NCOs. La frecuencia de SDSA es cercana al 90 %, con lo cual sólo el 10 % de los DSBs se resolverán en entrecruzamiento vía la DSBR.

mecansimos moleculares que llevan a cabo la recombinación (que incluye tanto al entrecruzamiento como a la conversión génica). En los próximos años, probablemente veremos propuestas de nuevos modelos que explicarán de manera más precisa estos mecanismos y poco a poco, podrán incorporarse a simulaciones evolutivas.

De hecho, en los últimos años se han descubierto una cantidad enorme de fenómenos evolutivos antes desconocidos. Por ejemplo, la conversión génica también se lleva a cabo entre copias duplicadas de genes o fragmentos de genes en un mismo cromosoma (interlocus) dando lugar a un fenómeno conocido como evolución concertada (la evolución no independiente de fragmentos de alta similitud en el genoma). Además, si llegara a ocurrir un *mismatch* (que un nucleótido no quede propiamente emparejado con su posición homóloga) después de un evento

de conversión génica (cosa no poco común), el error se corrige mediante un mecanismo conocido como *nucleotide excision repair*. Para complicar aún más el panorama, esta corrección ocurre de forma sesgada pues se privilegian las uniones G-C a las A-T (*biased gene conversion*). Evidencia de estos mecanismos ha salido a la luz sobre todo a partir de la secuenciación de genomas completos, mediante la cual, se han podido llevar a cabo mediciones precisas del número de duplicaciones segmentales y del contenido GC[8] de múltiples genomas, entre muchas otras cosas.

2.5. *Hotspots* y la distribución no homogénea de la recombinación

A partir de la secuenciación de genomas completos, en particular de seres humanos, se ha podido reconstuir con alta calidad el mapa de tasas de recombinación del genoma humano. Se ha encontrado que la distribución de eventos de recombinación no es homogénea sino que las tasas de recombinación varían en varios órdenes de magnitud (de $4x10^{-4}$ cM a 0.14 cM)[9] a lo largo del genoma [3].

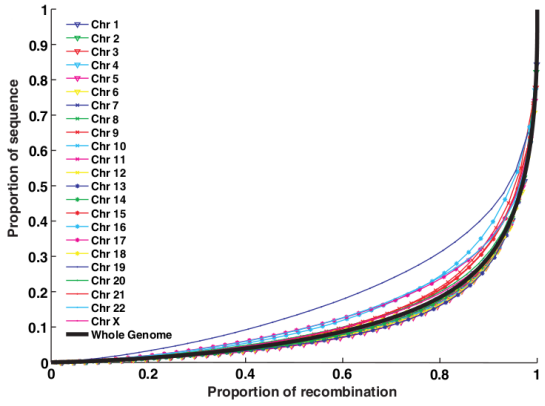

Figura 2.3: Tomada de [27]. Se muestra la proporción de la recombinación total que ocurre en varios porcentajes de la secuencia en cada cromosoma. Típicamente, el 80 % de la recombinación ocurre entre el 10 y 20 % de la secuencia. Una excepción interesante es el caso del cromosoma 19, que tiene una densidad e intensidad mucho menor de *hotspots*, relacionado seguramente con el hecho de que tiene la mayor densidad de genes [27].

[8] El contenido GC o *GC content* es la proporción local de pares de bases GC respecto a pares AT.

[9] Un centimorgan (cM) es la unidad estándar para medir la frecuencia de recombinacación. Dos pares de bases se encuentran a 1 cM de distancia si existe una probabilidad del 1 % de que entre ellos ocurra un entrecruzamiento en una generación. Es común que se utilice esta medida para referirse a distancias a lo largo del genoma, pero éstas varían dependiendo de las tasas de recombinación locales. En humanos, 1 centimorgan en promedio corresponde a una distancia de aproximadamente $7.5x10^5$ pares de bases [32].

A las regiones de los cromosomas donde los eventos de recombinación ocurren de manera más frecuente se les llama *hotspots* (o "puntos calientes [de recombinación]") y suelen ser de longitudes de una a dos kilobases de DNA [14]. Se han identificado más de 25000 probables *hotspots* en humanos, con uno aproximadamente cada 50 kb. Típicamente, el 80 % de los eventos de recombinación ocurren en el 10-20 % de la secuencia [27] (ver Figura 2.3).

Hotspots de 1-2 kb también se han caracterizado en la levadura (*Saccharomyces cerevisiae*) y en ratones, y la heterogeneidad en las tasas de recombinación también se ha observado en el maíz y en *Arabidopsis thaliana* [3]. Sin embargo, existen otros organismos modelo, como el gusano *Caenorhabditis elegans* y la mosca de la fruta *Drosophila melanogaster* en donde no se ha detectado evidencia de la presencia de *hotspots* [14].

Por otro lado, las tasas de recombinación están fuertemente correlacionadas (positivamente) con el contenido GC y con otras características, como la densidad de genes [17]. Sin embargo, al compararse regiones génicas[10] en humanos con regiones no génicas encontramos que la tasa de recombinación en las primeras es mucho menor que en las segundas, a razón de 0.75, con un peso estadístico significativo. Como se muestra en la figura 2.4, tomada de [23], las tasas de recombinación tienen una fuerte correlación negativa con las zonas donde se encuentran los genes. La aparente contradicción se explica si los *hotspots* de recombinación se encuentran con mayor probabilidad cerca de genes, pero no adentro de genes (ver figura 2.5).

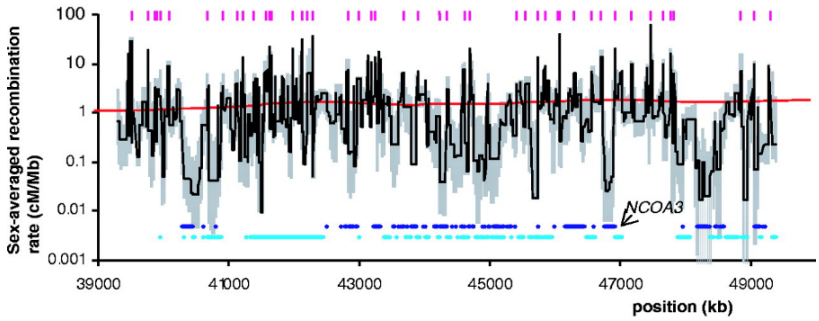

Figura 2.4: Tomada de [23]. Se muestran las tasas de recombinación estimadas para una región de 10 Mb en el cromosoma 20 de un individuo europeo (UK Caucasian) (línea negra). Las pequeñas líneas verticales rojas muestran la ubicación de los *hotspots* de recombinación con alto soporte estadístico y las líneas azules muestran la ubicación de los genes en las dos cadenas (positiva, azul obscuro; negativa, azul claro). Nótese la notable falta de recombinación en la zona del gen *NCOA3*.

Los *hotspots* tienden a concentrarse cerca de regiones promotoras pero a escala fina evitan secuencias directamente involucradas en funciones promotoras y regiones transcritas [23]. A partir de los datos obtenidos de humanos, se cree que los *hotspots* son bastante comunes y que están esparcidos por el genoma,

[10] Una región génica se define como la secuencia entre el inicio del primer exón y el final del último exón[23].

pero no parecen estar asociados a ninguna otra función, como sitios de anclaje de factores de transcripción [22]. Además y para resaltar su relevancia para este trabajo, parece que los *hotspots* evitan ubicarse adentro de genes. A pesar de que la evidencia que tenemos de este hecho proviene de humanos, es de esperarse que esté conservada en los organismos que tienen cromosomas diploides y posiblemente en muchos otros.

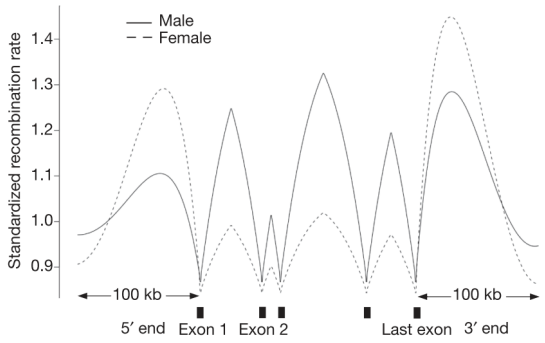

Figura 2.5: Tomada de [18]. Figura esquemática que representa las tasas de recombinación al interior de un gen. La tasa de recombinación es marcadamente baja en exones pero alta en intrones en zonas alejadas de los exones. Las tasas de recombinación suelen ser altas en regiones intergénicas a 40 kb del primer y el último exón y mayor en hembras que en machos.

2.6. La hipótesis de los bloques constructores

La idea de "bloques constructores" es súmamente intuitiva. El ejemplo típico que se usa para ilustrarlo, propuesto por Herbert Simon[33], es pensar en un relojero armando un reloj. Una primera posibilidad en que podemos imaginarlo es que vaya añadiendo pieza por pieza al reloj. Si se equivoca al colocar una pieza, deberá empezar todo desde cero. Otra posibilidad es que el relojero construya pequeños módulos funcionales uno por uno y que al final ensamble los módulos para armar el reloj. Esta segunda manera parece intuitivamente mejor, pues si se equivoca al armar un módulo, sólo tiene que empezar de nuevo ese módulo y no todo el reloj; si se equivoca ensamblando los módulos, sólo tiene que empezar la segunda etapa de nuevo, pero los módulos ya los tiene armados.

En los algoritmos genéticos, el concepto de bloques constructores se ha formalizado a partir del Teorema de Esquemas de Holland [15] (HST, por *Holland's Schema Theorem*) propuesto en 1975. El HST establece que en un algoritmo genético (que evoluciona mediante selección proporcional y entrecruzamiento en 1 punto), esquemas cortos, de bajo orden y con una adecuación por encima del promedio se incrementan de forma exponencial en generaciones sucesivas. Además, en los algoritmos genéticos opera la recombinación, y ésta permite tomar soluciones "parciales" de una parte de la población y combinarlas con soluciones "parciales" de otra parte de la población para generar soluciones

"completas". Así, el HST se complementa con la Hipótesis de Bloques Constructores [11] (BBH). Ésta última dice que las soluciones "parciales", que son precisamente los esquemas cortos, de bajo orden y con alta adecuación, y a los que llama *bloques constructores*, se recombinan de tal forma que permiten al algoritmo genético encontrar soluciones óptimas o cercanas a óptimas (*i.e.* cadenas con alta adecuación).

Desafortunadamente, no existe una caracterización de qué tan adecuados (de adecuación, como traducción de *fitness*), qué tan cortos y de qué tan bajo orden deben ser estos esquemas para poderlos considerar como bloques constructores [35], es decir, que por muy intuitiva que pueda ser la idea de bloques constructores, su representación real no es para nada evidente. Otro "punto débil" de la BBH es que de la misma manera en que la recombinación puede combinar soluciones parciales cortas para generar otras mejores y más largas, también puede destruir las soluciones óptimas ya creadas. De hecho, durante mucho tiempo todo el marco matemático existente en torno a la recombinación se basaba exclusivamente en la disrupción[11] de esquemas. A partir de 1999, se han dado a conocer teoremas exactos describiendo la creación de esquemas [37] y estableciendo criterios matemáticos para distinguir recombinación constructiva de recombinación disruptiva ([29], [36]).

Un ejemplo de la física que cumple con la hipótesis de bloques constructores es la nucleosíntesis estelar. La formación de átomos pesados se da a partir de una construcción jerárquica de soluciones parciales. Así, para formar un átomo de oxígeno, es mucho más sencillo primero construir helio a partir de dos hidrógenos, luego carbono a partir del helio, y finalmente oxígeno a partir del cabono.

En términos probabilísticos es mucho más sencillo ir construyendo átomos estables poco a poco hasta construir un átomo de 16 nucleones que unir en un solo evento 16 nucleones. Lozano, et al. [19], mostraron la existencia de una jeraquía de módulos y operadores que incrementaban la eficiencia del proceso de búsqueda de óptimos complicados, apoyando la BBH.

[11] Disrupción se utiliza como sinónimo de ruptura. A pesar de que la palabra "disrupción" aún no es aceptada por la Real Academia de la Lengua Española, se utiliza ampliamente desde algún tiempo y ya está registrada en el banco de datos de la RAE con lo cual es de esperarse que se acepte pronto.

Capítulo 3

Teoría de la selección y de la recombinación

"But if variations useful to any organic being do occur, assuredly individuals thus characterised will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterised. This principle of preservation, I have called, for the sake of brevity, Natural Selection."[5]

Charles Darwin

3.1. La selección

En 1859, Darwin propuso su teoría de evolución por selección natural. La selección natural se puede definir como el principio mediante el cual las pequeñas variaciones de rasgos que brinden a un individuo una mayor posibilidad de perseverar en la lucha por la supervivencia, tenderán a preservarse.

Posterior a la publicación de *El origen de las especies*, Herbert Spencer, equiparó a la selección natural de Darwin, con el concepto de la supervivencia del más adecuado[1] (*survival of the fittest*). Este término agradó a Darwin de tal forma que decidió incorporarlo a su texto en posteriores ediciones diciendo que la nueva expresión eliminaba la connotación atropocéntrica de la palabra "selección". Esta nueva expresión ha tenido muchos problemas (sobre todo por los otros usos que se le dan a la palabra *fitness* en inglés) pero hace evidente la necesidad de que exista un *fitness* asociado a cada individuo, para que se elijan a

[1]La traducción que generalmente se le da a *survival of the fittest* es "la supervivencia del más apto". Aquí se decidió utilizar la palabra "adecuado" para ser consistentes con la traducción de *fitness* como "adecuación".

los más adecuados. Sin el concepto de adecuación no tiene mucho sentido pensar en selección (a menos que ésta fuera aleatoria).

En el contexto de los algoritmos genéticos, imaginemos una población que evoluciona mediante mutación y recombinación, pero sin selección. Esta población podría explorar el espacio de búsqueda, pero de nada le serviría encontrar una solución óptima, pues pronto se perdería por la mutación y recombinación mismas. Con selección, se le da una razón de ser a la búsqueda de una solución óptima, pues una vez que ésta se encuentra, se "aprovecha". En este sentido, me permito afirmar que la selección es el operador fundamental de la evolución. La solución óptima a la que me refiero aquí no tiene que ser única. Sencillamente es una mejor solución que una elegida aleatoriamente, y de nuevo esto nos remite a la necesidad de que existe una medida de la adecuación de cada solución, para que tenga sentido la selección.

En cómputo evolutivo existen múltiples operadores de selección estándar. El más conocido, por su simplicidad, y porque es, en mi opinión, el más afín a lo planteado por Darwin, es la *selección proporcional*, que implementa la idea de que la probabilidad de selección es proporcional al valor de la adecuación del individuo en cuestión.

Sean $P_I(t)$ la fracción de individuos de una población representados por el esquema I, f_I la adecuación asociada al esquema I y $\overline{f}(t)$ la adecuación promedio de la población al tiempo t. La fracción de individuos de la población representados por el esquema I al tiempo $t+1$ bajo un esquema de selección proporcional será:

$$P_I(t+1) = \frac{f_I}{\overline{f}(t)} P_I(t) \tag{3.1}$$

Si la adecuación del esquema I es mayor a la adecuación promedio, el porcentaje de individuos de la población representados por el esquema I aumentará. De lo contrario, disminuirá. Bajo ciertas condiciones, se pueden encontrar soluciones exactas a esta ecuación y a otras ecuaciones asociadas a otros tipos de selección[40].

Cabe mencionar que el esquema de selección que se utilizará en las simulaciones computacionales mostradas en el capítulo 6 de esta tesis será *selección por torneo* que se da mediante la elección aleatoria de dos individuos de la población que "compiten" entre ellos de tal forma que el de mayor adecuación permanece[2] en la población y el otro no (a menos de que vuelva a elegirse y "gane" la competencia).

Es de esperarse que haber utilizado selección proporcional habría dado resultados cualitativa y cuantitativamente similares, pues ambos operadores de selección dependen fuertemente del paisaje de adecuación sobre el que actúan.

[2] En lugar de decir que "permanece", lo más correcto sería decir que el de mayor adecuación "es elegido para recombinarse con otro elegido de la misma forma".

3.2. Paisajes de adecuación

En un paisaje de adecuación a cada genotipo se le asocia una adecuación. En esta sección se definirán los paisajes de adecuación que se utilizarán a lo largo de este trabajo. Como se comentó en la sección 1.2.3, cada individuo de la población estará conformado por varios bloques y los paisajes de adecuación estarán asociados a los bloques. Todos los bloques de un individuo serán equivalentes, así como todos los individuos de la población. La adecuación total de cada inividuo será la suma de las adecuaciones de cada uno de los bloques que lo conforman.

3.2.1. Contando Unos

Uno de los paisajes de adecuación estándar en los algoritmos genéticos es el paisaje conocido como *Contando unos* (CU)[3]. Como su nombre lo indica, la adecuación de cada esquema dependerá del número de unos que éste contenga. Por ejemplo, el esquema 101, tendrá una adecuación de 2, mientras que el 111 tendrá una adecuación de 3. Como se muestra en la figura 3.1, si acomodamos los esquemas de forma adecuada queda en evidencia que para alcanzar el esquema óptimo (en este caso el 111), se puede ir "subiendo la escalera de adecuación". Los bloques constructores son explícitos en el sentido de que se favorece la construcción de los esquemas aptos poco a poco, es decir, la contrucción de esquemas de aquéllos de más baja adecuación (por ejemplo, el 010, o el 001) a aquéllos de más alta adecuación (por ejemplo, el 011).

3.2.2. Aguja en un pajar

Otro de los paisajes de adecuación estándar es el de *Aguja en un pajar* (AP)[4]. Como su nombre lo indica, entre todos los genotipos posibles existe sólo uno (la aguja) con una adecuación mayor a la del resto.

Elijamos un esquema, por simplicidad, el esquema de puros unos (111, para $l = 3$), y establezcamos que su adecuación sea 2. El esquema 111 será la aguja y todos los demás esquemas tendrán adecuación 1 y serán la paja. En este caso no existen los bloques constructores de forma explícita como con el paisaje CU. En el paisaje AP, para encontrar la aguja hay que recombinar dos esquemas que formen la aguja y esperar que no desaparezca, ya sea por deriva o por una posterior recombinación. Entre mayor sea l, más difícil será encontrar la aguja.

Desde el punto de vista biológico, no tiene mucho sentido pensar en un paisaje de adecuación como el AP. Es altamente improbable que para un gen sólo exista una configuración más apta que el promedio. Las variaciones que se le pueden hacer a este paisaje son infinitas. La pregunta complicada es cuáles de estos paisajes nos darán resultados diferentes al AP. Una modificiación al paisaje AP que resulta obvia es que en lugar de que toda la paja tenga una adecuación homogénea, ésta sea aleatoria dentro de un rango (de 0.5 a 1.5, por ejemplo).

[3] En la literatura, se utiliza comúnmente el término CO, por *Counting Ones*

[4] En la literatura, se utiliza comúnmente el término NIAH, por *Needle In A Haystack*

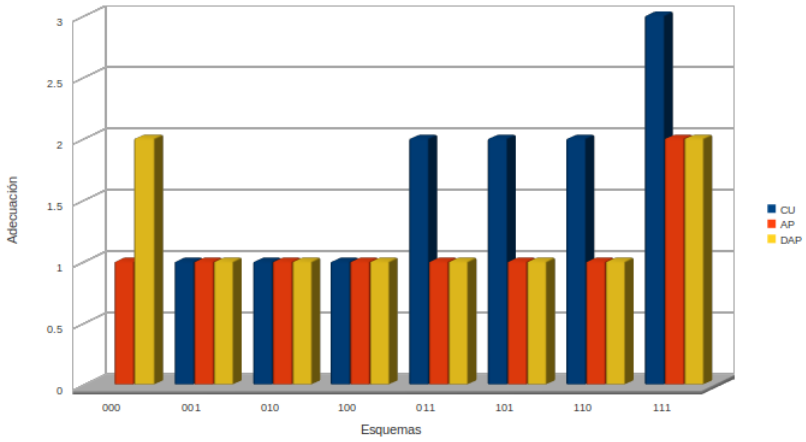

Figura 3.1: Adecuación de esquemas de $l = 3$ en tres paisajes de adecuación distintos: *Contando unos* (CU), *Aguja en un pajar* (AP) y *Doble aguja en un pajar* (DAP). El orden en que aparecen los esquemas no corresponde al orden habitual (i.e. por su conversión decimal), sino que están acomodados con el fin de mostrar más claramente la "escalera de adecuación" en el paisaje CU.

Esta pequeña variación sin duda lo hace más real, pero cualitativamente no existen diferencias sustanciales entre este paisaje y el AP estándar (datos no mostrados).

Otra posibilidad aún más realista desde el punto de vista biológico[5] es que en lugar de que exista sólo una configuración apta, existan varias que sean distintivamente más aptas que el promedio. Sabemos que existen múltiples genes en la naturaleza que tienen dos alelos[6]. Podemos considerar un paisaje de aduecuación tipo AP, pero que en lugar de tener una sola aguja, tenga dos. A este paisaje le llamaremos *Doble Aguja en un Pajar* (DAP). Para construir este paisaje hay que antes tomar en cuenta algunos detalles que se abordarán en la siguiente sección.

3.2.3. Doble aguja en un pajar

Imaginemos que tenemos un gen para el cual existen dos alelos identificados en la naturaleza. Sabemos que los alelos por definición deben ser diferentes fenotípicamente, y por lo tanto, serán también diferentes genotípicamente. La "distancia fenotípica" es difícil de calcular cualitativamente, sin embargo, en teoría debería ser relativamente sencillo calcular la "distancia genotípica" entre ambos alelos. Una posibilidad para calcular la distancia genotípica, sería por ejemplo, calcular la diferencia que existe entre las secuencias nucleotídicas

[5] A pesar de que se diga "aún más realista", con ello no queremos decir que en efecto sea "realista"; simplemente que es más realista que el AP estándar.

[6] Cuando se dice que un gen tiene x alelos, no implica que sólo existan x configuraciones viables. Tan sólo quiere decir que de todas las configuraciones posibles, sólo x se observan en la naturaleza.

Aguja	111110	110111	001111	011110	000111	010101	000000
111111	1 / 0	1 / 0	2 / 1	2 / 5	3 / 2	3 / 4	6 / 5

Cuadro 3.1: Distancia Hamming y distancia PPCD entre cada aguja y la aguja 111111.

correspondientes a cada uno de los dos alelos en cuestión. Sabemos que estas secuencias pueden ser redundantes, es decir, que dos secuencias sean diferentes pero que representen exactamente al mismo alelo (por la degeneración en la tercera base, por ejemplo). Siendo así, podemos contar el número de nucleótidos que son "necesariamente diferentes" entre dos secuencias nucleotídicas correspondientes a dos alelos de un mismo gen.

¿Será grande esta distancia o corta? Sabemos que existen alelos que son diferentes en únicamente una base nucleotídica, pero, ¿es esto una constante o la excepción en la naturaleza?

Para contruir el paisaje DAP estas preguntas son bastante relevantes pues *a priori* podemos suponer que no tendrá el mismo resultado utilizar un paisaje de adecuación que tenga dos agujas cuya distancia Hamming[7] sea igual a 1 que utilizar dos agujas cuya distancia Hamming sea igual a l.

Podemos calcular fácilmente la distancia Hamming entre cada posible par de agujas, sin embargo, quizás ésta no sea el mejor criterio para distinguir entre distintos paisajes DAP.

Pensando en bloques constructores y que en un momento dado las dos agujas podrían estarse segregando[8] en la población, tiene sentido considerar una posible recombinación entre ambas agujas. Dependiendo del punto de cruza elegido, la recombinación podrá no tener ningún efecto sobre ellas o "destrozarlas". Se puede entonces calcular el número de puntos de cruza que destrozan a las agujas en un evento de recombinación y a esta otra forma de medir la diferencia entre dos agujas le llamaremos *distancia por puntos de cruza destructivos* o *distancia PPCD*. Si recombinamos, por ejemplo, la aguja 111111 con la aguja 000000, sin importar el punto de cruza que se elija, ambas cadenas se destruirán y se obtendrá, por ejemplo, 110000 y 001111 si se recombina en el segundo punto de cruza. Por lo tanto, la distancia PPCD entre ambas agujas será 5. Si se recombina la aguja 111111 con la aguja 110011 sólo existe un punto de cruza (el del medio) que las destroza y forma 110111 y 111011, por lo tanto, la distancia PPCD entre ambas será 1. La distancia PPCD entre dos agujas siempre es menor a l pues para todo par de cadenas de igual longitud siempre existe un punto de cruza entre ellas menor a la longitud de las mismas. El cuadro 3.1 compara la distancia Hamming con la distancia PPCD entre distintos pares de agujas.

Para construir el paisaje DAP deberemos elegir a las agujas y asignarles a las secuencias correspondientes una adecuación mayor a la del resto. Dependiendo

[7]La distancia Hamming entre dos secuencias binarias es el número de diferencias por posición. Por ejemplo, la distancia Hamming entre 111 por un lado y 110, 101 ó 011 por otro, es uno, mientras que la distancia Hamming entre 101 y 010 ó entre 111 y 000 es tres.

[8]Se dice que un polimorfismo o un alelo se está segregando en una población cuando su frecuencia está entre 0 y 1, es decir, que se encuentra en algunos individuos de la población.

de las agujas elegidas se obtendrán resultados distintos. Ejemplos concretos de las diferencias entre la clasificación por distancia PPCD y por distancia Hamming se muestran en la sección 6.3.

3.3. Paisajes modulares

Una de las hipótesis principales de este trabajo es que la recombinación ha evolucionado a la par de la modularidad en el genoma. Así, la modularidad podría ser una consecuencia de la manera en que los genes se formaron en un principio. Nuestra hipótesis es que los genes se han ido construyendo poco a poco, mediante la unión sucesiva de módulos funcionales. Esta idea es en muchos sentidos similar a la expuesta por Holland [15] acerca de los bloques constructores (ver sección 2.6).

Si pensamos en genes modulares, debemos pensar en paisajes de adecuación modulares. Imaginemos que para formar un gen, requerimos que se combinen módulos aptos en una misma cadena. Supongamos que queremos constriur con gen con 4 módulos[9]. Ya se introdujeron en las secciones anteriores algunos paisajes de adecuación como el AP y el DAP. Utilizaremos precisamente estos paisajes de adecuación para calcular la adecuación de los módulos y la adecuación total del gen será la suma de las adecuaciones de los módulos que lo conforman.

De esta manera tenemos dos procesos simultáneos que se están llevando a cabo. Por un lado, la exploración de los paisajes de adecuación modulares para encontrar las secuencias más aptas, y por otro, la recombinación de estos módulos (más o menos aptos) para conformar los genes.

Es aquí donde se hace evidente que se necesitará definir el criterio mediante el cual consideraremos que una población es mejor o peor que otra, o que un modelo de recombinación es mejor o peor que otro.

3.4. Desequilibrio de ligamiento

En la genética de poblaciones existen múltiples métodos para detectar si una población ha estado sujeta, por ejemplo, a selección positiva, negativa o balanceadora, o si sufrió algún evento demográfico como una expansión poblacional o un cuello de botella reciente. También existen métodos para detectar ciertos patrones de recombinación. Un medida sencilla para determinar la epistasis entre dos loci es el desequilibrio de ligamiento, medida que también se ha adoptado y adaptado en los algoritmos genéticos. A continuación se describirán los conceptos de epistasis y desequilibrio de ligamiento para luego dar paso al desequilibrio de ligamiento determinado por selección que se utiliza en los algoritmos genéticos.

[9] Como en esta tesis sólo se tratarán cadenas de longitud fija, deberemos fijar la longitud de la cadena y la longitud de los módulos (y en consecuencia, el número de módulos por cadena) desde un principio.

3.4.1. Epistasis

El término *epistasis* se utiliza con significados ligeramente distintos dependiendo del contexto. En términos generales, se refiere al fenómeno mediante el cual los efectos de un gen se ven modificados por otro u otros genes. Estos efectos pueden ser fenotípicos, cuando el efecto de un gen se ve enmascarado por el efecto de otro (por ejemplo, el gen que causa albinismo enmascararía a los genes que determinan el color del pelo de una persona), aunque generalmente se emplea para referirse a efectos sobre la adecuación del gen en cuestión. En genética de poblaciones se consideran diferentes tipos de epistasis entre genes, dependiendo de la naturaleza estadística de las contribuciones de los genes a la adecuación de los individuos. Por ejemplo, una epistasis aditiva entre dos genes implica que las contribuciones de ambos genes a la adecuación del individuo son independientes; una epistasis multiplicativa entre dos genes implica que existe una correlación entre dichas contribuciones.

La epistasis y la interacción genética se refieren a diferentes aspectos de un mismo fenómeno. El término epistasis descrito anteriormente no implica necesariamente que exista una interacción bioquímica entre dos genes (por ejemplo, que la activación de un gen inhiba la expresión de otro); simplemente se refiere a una desviación de un comportamiento "independiente" del gen (epistasis multiplicativa contra epistasis aditiva, por ejemplo).

En nuestro contexto podemos definir el tipo de epistasis entre cualquiesquiera dos elementos de una cadena. Es decir, que además de determinar el tipo de epistasis entre dos bloques de una cadena (equivalente a epistasis entre genes), podemos también determinar, por ejemplo, el tipo de epistasis entre dos bits de un mismo bloque. En general, la epistasis entre todos los elementos de una cadena quedarán establecidos a partir del paisaje de adecuación sobre el cual estén evolucionando. El paisaje *Contando unos*, por ejemplo, establece una epistasis aditiva entre los bits de un mismo bloque, en el sentido de que la adecuación total del bloque es precisamente la suma de las contribuciones de cada bit (cada 1 en la cadena aumenta la adecuación en 1, independientemente de su ubicación en el bloque y del estado de sus bits vecinos[10]). A la epistasis aditiva también se le considera como ausencia de epistasis como consecuencia natural de que los elementos se comportan de forma independiente sin *interacción epistática*. En cambio, el paisaje *Aguja en un pajar* establece una epistasis máxima entre todos los bits de un mismo bloque (la presencia de un 1 en una posición es irrelevante a menos que haya un 1 en todas las demás posiciones del mismo bloque).

Por otro lado, cuando hablamos de paisajes de adecuación modulares estamos definiendo también el tipo de interacción epistática entre los bloques, que *por default*, digamos, es epistasis aditiva, es decir, con independencia entre bloques. En otras palabras, la adecuación de un individuo siempre será la suma de las adecuaciones de los bloques (o módulos) que lo conforman. Se podrían evidentemente, establecer diferentes tipos de epistasis entre bloques, pero para estudiar la recombinación asociada a la modularidad tiene sentido comenzar por el caso más sencillo que es epistasis aditiva entre bloques.

[10] Tomando como referencia los paisajes de la figura 3.1

La epistasis[11] es una de las causas de que exista *desequilibrio de ligamiento*.

3.4.2. El desequilibrio de ligamiento en la genética de poblaciones

En genética de poblaciones, se usa el desequilibrio de ligamiento para describir la asociación (o medir la epistasis) entre alelos en cromosomas. Consideremos una población de una especie diploide con dos loci ligados y cada locus con dos alelos segregantes (que aparecen en la población). Si estudiamos la dinámica poblacional con apareamiento aleatorio y recombinación veremos que el desequilibrio de ligamiento es un concepto que aparece naturalmente.

Sean A_1 y A_2 los dos alelos del primer locus y B_1 y B_2 los dos alelos del segundo locus, de tal manera que tengamos 4 gametos posibles en la población, A_1B_1, A_1B_2, A_2B_1 y A_2B_2 y sus respectivas frecuencias x_1, x_2, x_3 y x_4. La frecuencia del alelo A_1, como función de las frecuencias gaméticas es $p_1 = x_1+x_2$. Análogamente, la frecuencia del alelo B_1 es $p_2 = x_1 + x_3$[12].

Sea r la tasa de recombinación. Después de una generación, es decir, después de apareamiento aleatorio y recombinación con probabilidad r entre los individuos de la población, la frecuencia del gamento A_1B_1 será

$$x_1' = (1 - r)x_1 + rp_1p_2. \tag{3.2}$$

Analicemos rápidamente esta expresión. La probabilidad de elegir a un gameto A_1B_1 es proporcional a su frecuencia x_1. Una vez elegido, tendrá dos posibilidades: que no se recombine (esto ocurre con probabilidad $1 - r$), con lo cual permanecerá en la población como A_1B_1 o que sí recombine (que ocurre con probabilidad r), con lo cual dejará de ser A_1B_1. La otra posibilidad de tener un gameto A_1B_1 en la población es si resulta de una recombinación de los alelos A_1 y B_1 y esto ocurrirá con probabilidad rp_1p_2. Las frecuencias gaméticas p_1p_2 pueden multiplicarse pues el apareamiento aleatorio implica que la selección de alelos en cada locus es independiente.

Por lo tanto, el cambio en la frecuencia del gameto A_1B_1 después de una generación de apareamiento aleatorio será

$$x_1' - x_1 = -r(x_1 - p_1p_2). \tag{3.3}$$

Definimos el desequilibrio de ligamiento como

$$\Delta = x_1 - p_1p_2, \tag{3.4}$$

[11] Formalmente se debería hacer la diferenciación entre epistasis en general y epistasis con efectos sobre la adecuación (en inglés, *fitness epistasis*). En este trabajo, toda la epistasis que tratemos será epistasis con efectos sobre la adecuación, pero por simplicidad, se utilizará simplemente el término epistasis.

[12] Ésta es la notación estándar en la genética de poblaciones para denotar las frecuencias gaméticas. Se decidió conservar esta notación en este capítulo para diferenciarla claramente del desequilibrio de ligamiento determinado por selección, que es el que usaremos más adelante.

que es precisamente una medida de la diferencia entre la frecuencia real del gameto A_1B_1 (en este caso) y el valor esperado de esta frecuencia si la asociación de alelos ocurriera de manera totalmente aleatoria. Si en la población observamos que la frecuencia real difiere de la frecuencia esperada (de forma estadísticamente significativa), quiere decir que existe una razón por la cual el alelo A_1 se asocia preferencialemente con el alelo B_1. Esto se puede deber a diversos motivos, pero el más sencillo es que exista una presión selectiva que favorezca esta asociación.

Si Δ es positiva implica que la selección natural favorece el acoplamiento entre los gametos A_1 y B_1; si es negativa, implica que el acoplamiento es desfavorable. Por lo tanto, una Δ significativamente distinta de cero implicará probablemente que existe epistasis, positiva o negativa, entre los loci en cuestión.

3.4.3. Desequilibrio de ligamiento determinado por selección en los algoritmos genéticos

El paso del desequilibrio de ligamiento, a secas, al desequilibrio de ligamiento determinado por selección es muy representativo del paso de la evolución *in vivo* a la evolución *in silico*[13]. Como se comentó en la sección anterior, el desequilibrio de ligamiento distinto de cero "puede deberse a diversos motivos, pero el más sencillo es que exista una presión selectiva que favorezca la asociación". En una simulación computacional, los "diversos motivos" pueden reducirse (si se quiere, y como se hace generalmente) al "más sencillo". La "presión selectiva" se traduce por lo tanto a que los esquemas en cuestión tienen una adecuación mayor al promedio. El desequilibrio de ligamiento determinado por selección, se calculará por lo tanto, y como indica su nombre a partir de las frecuencias $P_I^{'}(t)$.

Sin embargo, ésta no será la única diferencia respecto al desequilibrio de ligamiento presentado en la sección anterior. Como se comentó, en genética de poblaciones, el desequilibrio de ligamiento se utiliza para medir el grado de epistasis que existe entre dos loci. Ahora, en lugar de alelos en loci diferentes, tomaremos módulos en un mismo locus. Es decir, que no vamos a ver la diferencia entre la frecuencia real de un gameto y su frecuencia esperada por asociación aleatoria, sino la diferencia entre la frecuencia real de un gen y la frecuencia esperada de que los módulos que lo conforman se recombinen para formarlo. En la representación de esquemas, estos módulos son precisamente los bloques constructores que forman una cadena determinada.

En analogía a la ecuación 3.4 describiremos una ecuación para el desequilibrio de ligamiento determinado por selección tomando como base la de bloques constructores[14].

[13] El término *in silico* se utiliza para referirse a los experimentos realizados mediante simulaciones computacionales, en analogía a los términos *in vivo* e *in vitro* utilizados comunmente en biología. El término se utilizó por primera vez por Pedro Miramontes en 1989 [comunicación personal].

[14] La ecuación también se puede deducir de la ecuación 4.8 aplicándole un "granulado grueso" y una transformación de coordenadas e ignorando la mutación (ver [35]).

El desequilibrio de ligamiento determinado por selección será función del tiempo y de la máscara m utilizada para recombinar los esquemas ya existentes. Sean I_m y $I_{\overline{m}}$ esquemas conjugados (o complementarios) que al recombinarse con máscara m dan lugar a I.

Podemos definir el desequilibrio de ligamiento determinado por selección ("selection-weighted linkage disequilibrium") como

$$\Delta_I(m,t) = P_I^{'}(t) - P_{I_m}^{'}(t)P_{I_{\overline{m}}}^{'}(t) \tag{3.5}$$

donde $P_{I_m}^{'}(t)$ es la probabilidad de elegir al esquema I_m al tiempo t y $P_{I_{\overline{m}}}^{'}(t)$ es la probabilidad de elegir al esquema $I_{\overline{m}}$ al tiempo t. Estos esquemas conjugados están determinados tanto por la máscara m como por la cadena I. Si, por ejemplo, $I = 101$ y $m = 100$ implica que escogeremos al segundo y tercer bit del primer padre y al primer bit del segundo padre para conformar la cadena I. Por lo tanto, $I_m = *01$, $I_{\overline{m}} = 1**$ y $\Delta_{101}(100,t) = P_{101}^{'}(t) - P_{*01}^{'}(t)P_{1**}^{'}(t)$.

También podemos calcular, por ejemplo, $\Delta_{I_m}(m^{'},t)$. Siguiendo el ejemplo anterior, tomemos $I_m = *01$, $m^{'} = 110$ [15] y obtenemos $\Delta_{*01}(110,t) = P_{*01}^{'}(t) - P_{**1}^{'}(t)P_{*0*}^{'}(t)$.

Si consideramos, como hemos venido haciendo hasta ahora que la sobre la población actúa selección proporcional, tendremos:

$$P_I^{'}(t) = \frac{f_I}{\overline{f}(t)}P_I(t). \tag{3.6}$$

A juzgar por la ecuación 3.6, la selección de una cadena específica, ya sea para ser clonada o para ser recombinada con otra cadena, está determinada exclusivamente por su adecuación relativa a las demás. A diferencia de la ecuación 3.4, la ecuación 3.5 no sólo sirve para analizar si existe una presión selectiva que hace que dos alelos aparezcan juntos más de lo esperado, sino que establece *a priori* la utilidad de llevar a cabo la recombinación para una población dada, con una máscara y en una generación específicas y dependiendo del paisaje de adecuación que se esté utilizando.

$\Delta_I(m,t) < 0$ implica que la recombinación es favorable, es decir, que la recombinación utilizando la máscara m produce más cadenas del tipo I en la próxima generación de las que se tendrían por seleccionar las cadenas I ya existentes en la población. Por lo contrario, $\Delta_I(m,t) > 0$ implica que la recombinación es desfavorable; en otras palabras, que la recombinación con la máscara m destruye más cadenas del tipo I de las que construye.

Para evaluar los efectos de la recombinación es necesario hacerlo máscara por máscara, pues cada máscara puede tener un efecto distinto. Además, es evidente que la utilidad de la recombinación dependerá muy sensiblemente del paisaje de adecuación que se utilice.

Consideremos una población inicial homogénea de N individuos, de tal manera que $P_I = P_{I_m} = P_{I_{\overline{m}}} = \frac{1}{2^N}$. De ahí:

[15] Tomando $m^{'} = 010$ daría el mismo resultado que con $m^{'} = 110$, pues $I_m = *01$, sólo que no tiene sentido para entrecruzamiento en un solo punto.

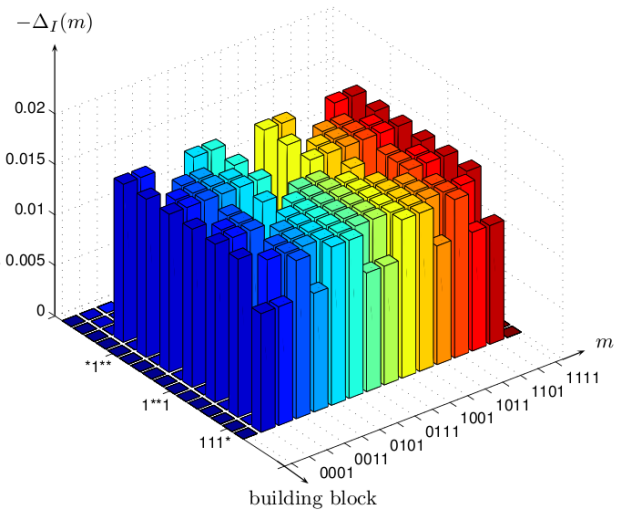

Figura 3.2: Tomada de [31]. Se muestra el resultado de calcular $-\Delta_I(m)$ para el paisaje CU en la representación de bloques constructores. La recombinación resulta siempre benéfica, es decir que la recombinación ayuda tanto a la construcción del genotipo óptimo como a la de los bloques constructores de dicha cadena. El desequilibrio de ligamiento para esquemas de orden 1 es neutro, así como para las máscaras donde todo un padre se elige como bloque constructor. Además, las máscaras que construyen al óptimo de forma simétrica, es decir, con 2 componentes de cada padre, resultan más benéficas que las no simétricas (3 y 1).

$$\begin{aligned}\Delta_I(m,t) &= P_I^{'}(t) - P_{I_m}^{'}(t)P_{I_{\overline{m}}}^{'}(t) \\ &= \frac{f_I}{\overline{f}(t)}P_I(t) - \frac{f_{I_m}}{\overline{f}(t)}P_{I_m}(t)\frac{f_{I_{\overline{m}}}}{\overline{f}(t)}P_{I_{\overline{m}}}(t) \\ &= \frac{1}{\overline{f}^2(t)2^N}\left(f_I(t)\overline{f}(t) - f_{I_m}(t)f_{I_{\overline{m}}}(t)\right)\end{aligned} \tag{3.7}$$

Recientemente, Rosenblueth y Stephens han calculado el desequilibrio de ligamiento para paisajes de adecuación sencillos como CU y AP ([30], [31]). El desequilibrio de ligamiento se puede calcular en diferentes bases y la óptima para mostrar sus efectos es precisamente la base de bloques constructores. Si consideramos cadenas donde cada bit representa un bloque constructor, la base de bloques constructores estará conformada por los esquemas que representen a todas las combinaciones posibles de bloques presentes o no en la cadena. Es decir, si nuestro bloque óptimo es el 1111, la base de bloques constructores estará conformada por los esquemas ****, ***1, **1*, **11, *1**, etc.

Las figuras 3.2 y 3.3 muestran los resultados de calcular el desequilibrio de ligamiento (ecuación 3.7) para cadenas de longitud 4 en paisajes CU y AP respectivamente, en la base de bloques constructores. Los resultados son claros y muestran que la recombinación es benéfica (o neutra) en el caso del paisaje CU

para todas los esquemas y todas las máscaras, mientras que en el paisaje AP, la recombinación siempre afecta a la población (o es neutra), independientemente de su estado y la máscara utilizada.

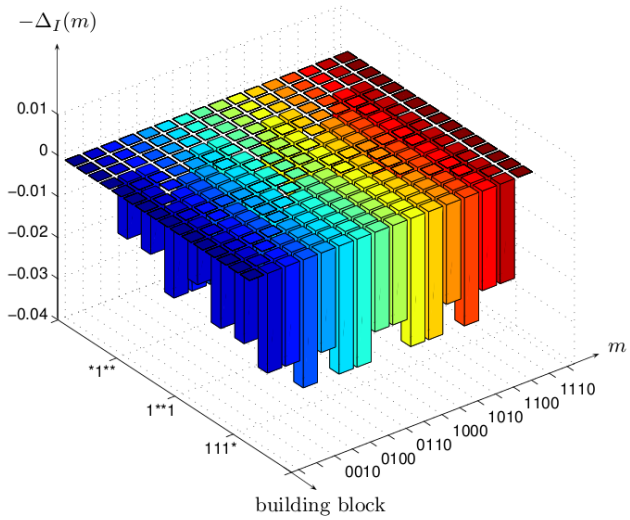

Figura 3.3: Tomada de [31]. Se muestra el resultado de calcular $-\Delta_I(m)$ para el paisaje AP en la representación de bloques constructores. Contrariamente a lo mostrado en la figura 3.2 para el paisaje CU, en el paisaje AP (de máxima epistasis), toda máscara lleva a un desequilibrio de ligamiento positivo para todo bloque constructor del genotipo óptimo.

3.5. La recombinación

La recombinación es otro de los operadores genéticos fundamentales. En esta tesis se tratará únicamente el caso de la recombinación en un solo punto (*one-point crossover*), es decir, que en cada evento de recombinación entre dos cadenas, éstas se entrecruzarán exclusivamente en un punto.

Ya se han descrito con cierto detalle los mecanismos involucrados en la recombición genética en la sección 2.4. A continuación, se explicarán los fundamentos matemáticos de la recombinación.

Para simplificar la explicación, tomaremos como ejemplo la recombinación en cadenas binarias de longitud 3, aunque esta teoría es generalizable a cadenas de cualquier longitud.

Si la longitud de nuestras cadenas es 3, tendremos 8 posibles estados de la cadena: 000, 001, 010, 011, 100, 101, 110 y 111. Como se explicó en **??**, podemos representar grupos de estas cadena mediante esquemas: por ejemplo, el esquema 1*1 representa las cadenas 101 y 111.

Consideremos una población de N individuos, cada uno de ellos representado por una cadena de 3 bits. Supongamos un modelo evolutivo sin selección, ni mu-

tación, sino únicamente recombinación. Para efectuar la recombinación se eligen 2 individuos al azar y se recombinan con probabilidad p_c en un punto aleatorio (para una cadena de 3 bits, hay 2 posibles sitios de recombinación). Este proceso se lleva a cabo $\frac{N}{2}$ veces cada generación, y así, p_c será la probabilidad de que el individuo i se recombine al tiempo t.

Sea $P_{111}(t)$ la fracción de individuos de la población al tiempo t representados por la cadena 111[16]. Análogamente, $P_{1*1}(t)$ será el porcentaje de individuos de la población representados por la cadena 111 o por la cadena 101.

Existen dos posibilidades para que un individuo de la población esté representado por la cadena 111 al tiempo t: que ya haya estado en la población al tiempo anterior y se haya elegido pero no recombinado, o que el individuo sea el resultado de una recombinación que dio lugar a la cadena 111.

Teniendo esto en cuenta podemos escribir, por ejemplo, $P_{111}(t+1)$ como función de $P_{111}(t)$ de la siguiente manera:

$$P_{111}(t+1) = (1-p_c)P_{111}(t) + \frac{p_c}{2}P_{11*}(t)P_{**1}(t) + \frac{p_c}{2}P_{1**}(t)P_{*11}(t) \tag{3.8}$$

El primer término de la suma corresponde al caso de no recombinación o clonación; el segundo y tercer términos de la suma corresponden a los dos posibles casos de recombinación que dan como resultado una cadena 111.

Análogamente, tenemos que

$$P_{110}(t+1) = (1-p_c)P_{110}(t) + \frac{p_c}{2}P_{11*}(t)P_{**0}(t) + \frac{p_c}{2}P_{1**}(t)P_{*10}(t) \tag{3.9}$$

Podemos ahora escribir $P_{11*}(t+1)$ como la suma de $P_{111}(t+1)$ y $P_{110}(t+1)$:

$$\begin{aligned} P_{11*}(t+1) &= P_{111}(t+1) + P_{110}(t+1) \\ &= (1-p_c)\left[P_{111}(t) + P_{110}(t)\right] + \frac{p_c}{2}P_{11*}(t)\left[P_{**1}(t) + P_{**0}(t)\right] \\ &\quad + \frac{p_c}{2}P_{1**}(t)\left[P_{*11}(t) + P_{*10}(t)\right] \\ &= (1-p_c)P_{11*}(t) + \frac{p_c}{2}P_{11*}(t) + \frac{p_c}{2}P_{1**}(t)P_{*1*}(t) \\ &= (1-\frac{p_c}{2})P_{11*}(t) + \frac{p_c}{2}P_{1**}(t)P_{*1*}(t) \end{aligned} \tag{3.10}$$

Análogamente,

$$P_{10*}(t+1) = (1-\frac{p_c}{2})P_{10*}(t) + \frac{p_c}{2}P_{1**}(t)P_{*0*}(t), \tag{3.11}$$

$$P_{*11}(t+1) = (1-\frac{p_c}{2})P_{*11}(t) + \frac{p_c}{2}P_{**1}(t)P_{*1*}(t),\ y \tag{3.12}$$

[16] $P_{111}(t)$ también puede interpretarse como la probabilidad de que si elegimos un individuo al azar de la población al tiempo t, éste esté representado por la cadena 111.

$$P_{*10}(t+1) = (1 - \frac{p_c}{2})P_{*10}(t) + \frac{p_c}{2}P_{**0}(t)P_{*1*}(t). \tag{3.13}$$

Por otro lado, podemos escribir P_{1**} como la suma de 3.10 y 3.11:

$$\begin{aligned} P_{1**}(t) &= P_{11*}(t) + P_{10*}(t) \\ &= (1 - \frac{p_c}{2}) \left[P_{11*}(t-1) + P_{10*}(t-1)\right] \\ &\quad + \frac{p_c}{2}P_{1**}(t-1) \left[P_{*1*}(t-1) + P_{*0*}(t-1)\right] \\ &= (1 - \frac{p_c}{2})P_{1**}(t-1) + \frac{p_c}{2}P_{1**}(t-1) \\ &= P_{1**}(t-1) \end{aligned}$$

Por lo tanto,

$$P_{1**}(t) = P_{1**}(0) \tag{3.14}$$

Es decir, que el número de esquemas de orden 1 son constantes en el tiempo. Siendo así, $P_{*1*}(t) = P_{*1*}(0)$ y al sustituir en la ecuación 3.10 obtenemos una ecuación lineal inhomogénea:

$$P_{11*}(t+1) = (1 - \frac{p_c}{2})P_{11*}(t) + \frac{p_c}{2}P_{1**}(0)P_{*1*}(0) \tag{3.15}$$

La ecuación 3.15 muestra como los bloques constructores de los esquemas de orden 2 (como 11*), son esquemas de orden 1 (como 1** y *1*).

Sean $(1 - \frac{p_c}{2}) = A$ y $\frac{p_c}{2}P_{1**}(t)P_{*1*}(t) = \frac{p_c}{2}P_{1**}(0)P_{*1*}(0) = C$. De la ecuación 3.10 obtenemos:

$$P_{11*}(t+1) = AP_{11*}(t) + C \tag{3.16}$$

Evaluemos 3.16 en $t = 0$, $t = 1$ y $t = 2$:

$$P_{11*}(1) = AP_{11*}(0) + C$$

$$\begin{aligned} P_{11*}(2) &= AP_{11*}(1) + C \\ &= A(AP_{11*}(0) + C) + C \\ &= A^2P_{11*}(0) + AC + C \end{aligned}$$

$$\begin{aligned} P_{11*}(3) &= A^3P_{11*}(0) + A^2C + AC + C \\ &= A^3P_{11*}(0) + C(A^2 + A + 1) \end{aligned}$$

Por inducción:

$$\begin{aligned} P_{11*}(t) &= A^tP_{11*}(0) + C(A^{t-1} + A^{t-2} + ... + A + 1) \\ &= \left(1 - \frac{p_c}{2}\right)^t P_{11*}(0) + \frac{p_c}{2}\left(\sum_{k=0}^{t-1}(1 - \frac{p_c}{2})^k\right)P_{1**}(0)P_{*1*}(0) \end{aligned} \tag{3.17}$$

Bajo el mismo razonamiento, y partiendo de la ecuación 3.12, tenemos que

$$P_{*11}(t) = \left(1 - \frac{p_c}{2}\right)^t P_{*11}(0) + \frac{p_c}{2}\left(\sum_{k=0}^{t-1}\left(1 - \frac{p_c}{2}\right)^k\right) P_{**1}(0)P_{*1*}(0) \tag{3.18}$$

Sabemos que si $x \geq 0$,

$$x\left(\sum_{k=0}^{n}(1-x)^k\right) = 1 - (1-x)^{n+1} \tag{3.19}$$

Por lo tanto,

$$\frac{p_c}{2}\left(\sum_{k=0}^{t-1}\left(1 - \frac{p_c}{2}\right)^k\right) = 1 - \left(1 - \frac{p_c}{2}\right)^t \tag{3.20}$$

Sustituyendo 3.20 en 3.17 obtenemos:

$$\begin{aligned} P_{11*}(t) &= \left(1 - \frac{p_c}{2}\right)^t P_{11*}(0) + \left(1 - \left(1 - \frac{p_c}{2}\right)^t\right) P_{1**}(0)P_{*1*}(0) \\ &= \left(1 - \frac{p_c}{2}\right)^t (P_{11*}(0) - P_{1**}(0)P_{*1*}(0)) + P_{1**}(0)P_{*1*}(0) \end{aligned}$$

Retomando la ecuación 3.5 donde se definió el desequilibrio de ligamiento, vemos que

$$\Delta_{11*}(0) = P_{11*}(0) - P_{1**}(0)P_{*1*}(0),$$

por lo tanto,

$$P_{11*}(t) = \left(1 - \frac{p_c}{2}\right)^t \Delta_{11*}(0) + P_{1**}(0)P_{*1*}(0) \tag{3.21}$$

Análogamente, de la ecuación 3.18 obtenemos:

$$P_{*11}(t) = \left(1 - \frac{p_c}{2}\right)^t \Delta_{*11}(0) + P_{**1}(0)P_{*1*}(0) \tag{3.22}$$

donde $\Delta_{*11}(0) = P_{*11}(0) - P_{**1}(0)P_{*1*}(0)$.

Analicemos esta ecuación. Por un lado, podemos verificar que si $p_c = 0$, $P_{*11}(t) = P_{*11}(0)$. Si $p_c \neq 0$, podemos ver el comportamiento cuando $t \to \infty$. Como $\lim_{t\to\infty}\left(1 - \frac{p_c}{2}\right)^t = 0$ entonces,

$$\lim_{t\to\infty} P_{*11}(t) = P_{**1}(0)P_{*1*}(0)$$

La dinámica de los esquemas de orden 2 está asintóticamente determinada por las frecuencias iniciales de los esquemas de orden 1. Aquí tenemos el primer ejemplo del "granulado grueso" que se comentó en la sección 1.3: para conocer

la dinámica de los esquemas de orden 2, nos estamos remitiendo a la dinámica de esquemas de orden 1, es decir, a un granulado más grueso.

Si sustituimos 3.21 y 3.22 en 3.8 obtenemos una ecuación recursiva lineal inhomogénea:

$$\begin{aligned}
P_{111}(t+1) &= (1-p_c)P_{111}(t) + \left(\frac{p_c}{2}\right)\left[\left(1-\frac{p_c}{2}\right)^t \Delta_{11*}(0) + P_{1**}(0)P_{*1*}(0)\right] P_{**1}(0) \\
&\quad + \left(\frac{p_c}{2}\right)\left[\left(1-\frac{p_c}{2}\right)^t \Delta_{*11}(0) + P_{**1}(0)P_{*1*}(0)\right] P_{1**}(0) \\
&= (1-p_c)P_{111}(t) + 2\left(\frac{p_c}{2}\right) P_{1**}(0)P_{*1*}(0)P_{**1}(0) \\
&\quad + \left(1-\frac{p_c}{2}\right)^t \left[\left(\frac{p_c}{2}\right)\Delta_{11*}(0)P_{**1}(0) + \left(\frac{p_c}{2}\right)\Delta_{*11}(0)P_{1**}(0)\right] \\
&= BP_{111}(t) + A^t\left[\left(\frac{p_c}{2}\right)\Delta_{11*}(0)P_{**1}(0) + \left(\frac{p_c}{2}\right)\Delta_{*11}(0)P_{1**}(0)\right] \\
&\quad + p_c P_{1**}(0)P_{*1*}(0)P_{**1}(0) \\
&= BP_{111}(t) + A^t E + F
\end{aligned} \tag{3.23}$$

donde $A = (1-\frac{p_c}{2})$, $B = (1-p_c)$, $F = p_c P_{1**}(0)P_{*1*}(0)P_{**1}(0)$ y $E = \left[\left(\frac{p_c}{2}\right)\Delta_{11*}(0)P_{**1}(0) + \left(\frac{p_c}{2}\right)\Delta_{*11}(0)P_{1**}(0)\right]$.

Evaluemos 3.23 en $t=0$, $t=1$, $t=2$ y $t=3$:

$$P_{111}(1) = BP_{111}(0) + E + F$$

$$\begin{aligned}
P_{111}(2) &= BP_{111}(1) + AE + F \\
&= B(BP_{111}(0) + E + F) + AE + F \\
&= B^2P_{111}(0) + E(B+A) + F(B+1)
\end{aligned}$$

$$\begin{aligned}
P_{111}(3) &= B(B^2P_{111}(0) + E(B+A) + F(B+1)) + A^2E + F \\
&= B^3P_{111}(0) + E(B^2+AB) + F(B^2+B) + A^2E + F \\
&= B^3P_{111}(0) + E(B^2+AB+A^2) + F(B^2+B+1)
\end{aligned}$$

$$\begin{aligned}
P_{111}(4) &= B(B^3P_{111}(0) + E(B^2+AB+A^2) + F(B^2+B+1)) + A^3E + F \\
&= B^4P_{111}(0) + E(B^3+AB^2+A^2B+A^3) + F(B^3+B^2+B+1)
\end{aligned}$$

Por inducción:

$$
\begin{aligned}
P_{111}(t) &= B^t P_{111}(0) + E(B^{t-1} + AB^{t-2} + ... + A^{t-2}B + A^{t-1}) \\
&\quad + F(B^{t-1} + B^{t-2} + ... + B + 1) \\
&= (1-p_c)^t P_{111}(0) + p_c \left(\sum_{k=0}^{t-1} (1-p_c)^k \right) P_{1**}(0) P_{*1*}(0) P_{**1}(0) \\
&\quad + \left(\frac{p_c}{2}\right) \sum_{k=0}^{t-1} (1-p_c)^{t-1-k} \left(1 - \frac{p_c}{2}\right)^k \left[\Delta_{11*}(0) P_{**1}(0) + \Delta_{*11}(0) P_{1**}(0)\right]
\end{aligned} \tag{3.24}
$$

Por la ecuación 3.19 sabemos que:

$$
p_c \left(\sum_{k=0}^{t-1} (1-p_c)^k \right) = 1 - (1-p_c)^t \tag{3.25}
$$

Además, se puede mostrar que:

$$
\left(\frac{p_c}{2}\right) \sum_{k=0}^{t-1} (1-p_c)^{t-1-k} \left(1 - \frac{p_c}{2}\right)^k = \left(1 - \frac{p_c}{2}\right)^t - (1-p_c)^t \tag{3.26}
$$

Sustituyendo 3.25 y 3.26 en 3.24 obtenemos:

$$
\begin{aligned}
P_{111}(t) &= (1-p_c)^t P_{111}(0) + \left(1 - (1-p_c)^t\right) P_{1**}(0) P_{*1*}(0) P_{**1}(0) \\
&\quad + \left[\left(1 - \frac{p_c}{2}\right)^t - (1-p_c)^t\right] \left[\Delta_{11*}(0) P_{**1}(0) + \Delta_{*11}(0) P_{1**}(0)\right] \\
&= (1-p_c)^t \left[P_{111}(0) - \Delta_{11*}(0) P_{**1}(0) - \Delta_{*11}(0) P_{1**}(0)\right. \\
&\quad \left. - P_{1**}(0) P_{*1*}(0) P_{**1}(0)\right] + \left(1 - \frac{p_c}{2}\right)^t \left[\Delta_{11*}(0) P_{**1}(0)\right. \\
&\quad \left. + \Delta_{*11}(0) P_{1**}(0)\right] + P_{1**}(0) P_{*1*}(0) P_{**1}(0)
\end{aligned} \tag{3.27}
$$

Si $p_c \neq 0$, $\lim_{t\to\infty} (1-p_c)^t = 0$ y $\lim_{t\to\infty} \left(1 - \frac{p_c}{2}\right)^t = 0$. Por lo tanto, si $p_c \neq 0$, entonces,

$$
\lim_{t\to\infty} P_{111}(t) = P_{1**}(0) P_{*1*}(0) P_{**1}(0) \tag{3.28}
$$

Tiene sentido que, en ausencia de selección, el número de individuos representados por un esquema, sea la multiplicación de las máscaras que dan lugar a ese esquema.

Análogamente, a las ecuaciones 3.14, 3.21 y 3.27, podemos encontrar soluciones exactas de $P(t)$ para cualquier esquema. En particular, encontramos soluciones exactas para las 2^l diferentes cadenas de l bits.

Originalmente pudimos haber pensado que para encontrar las soluciones exactas para cada cadena era necesario resolver 2^l ecuaciones homogéneas cuadráticas acopladas. Sin embargo, hemos visto que si resolvemos las ecuaciones de

manera jerárquica, es decir, comenzando por los esquemas de orden 1, luego los de orden 2 y así suscesivamente, sólo tendremos que resolver ecuaciones lineales inhomogéneas no acopladas. Éste es un claro ejemplo de cómo el "coarse graining" es sumamente útil para resolver ciertos problemas.

Capítulo 4

Modelos

Uno de los propósitos principales de esta tesis es mostar que tiene mucho sentido pensar que la recombinación modular tuvo efectos importantes en la conformación del genoma. Para ello, se analizarán algunos rasgos de la fenomenología de la recombinación adaptativa mediante simulaciones computacionales.

Antes de adentrarnos en ese terreno vale la pena presentar un modelo sencillo de la evolución del genoma para entender en términos generales de qué forma operan los principales operadores de la evolución, a saber, la selección, la recombinación (generalizada) y la mutación.

Las ecuaciones que se presentarán a continuación son generales, para cualquier tipo de selección, recombinación y mutación. En nuestras simulaciones elegiremos operadores específicos (selección por torneo, recombinación mediante entrecruzamiento en un solo punto y mutación nula). Dado que el propósito de este trabajo no es precisamente contrastar un modelo matemático con simulaciones computacionales, no se profundizará en la descripción matemática de los operadores específicos de las simulaciones. Es muy pertinente, sin embargo, presentar un modelo general que describa mediante la representación ya utilizada en el capítulo anterior, la evolución de una población.

Consideremos a una población de tamaño constante que evoluciona a lo largo del tiempo. Cada individuo de la población estará representado por un esquema I. En este caso, el esquema será binario y de longitud fija y en los ejemplos tomaremos $l = 3$ por simplicidad. $P_I(t)$ será la fracción de individuos de la población representados por el esquema I en el tiempo t.

Dado que el orden en el cual se aplican los operadores es muy relevante, presentaremos las ecuaciones pertinentes a cada operador en el mismo orden en el que se aplicarán, es decir, primero selección, luego recombinación y finalmente, mutación. La ecuación final describirá el cambio en las frecuencias genotípicas de la población.

4.1. Modelo evolutivo

4.1.1. Selección

Cuando sobre una población actúa el operador de selección, se elegirán de la población aquellos individuos que tengan una adecuación mayor mediante un mecanismo estocástico. Denotaremos como $P_I^{'}(t)$ a la fracción de individuos representados por el esquema I que permanecen en la población después de que sobre ella actúe el operador de selección elegido y antes de que actúen los demás operadores. La forma explícita de $P_I^{'}$ dependerá del tipo de selección que se utilice. Para el caso se selección proporcional (una de las selecciones estándar en algoritmos genéticos) y como se explicó en la sección 3.1:

$$P_I^{'}(t) = \frac{f_I}{\overline{f}(t)} P_I(t). \tag{4.1}$$

4.1.2. Recombinación

Después de aplicarse el operador de selección en toda la población, se aplicará el operador de recombinación. Describiremos el operador de selección de la forma más general posible, aunque como se comentará, nuestro modelo estará limitado a entrecruzamiento de cadenas en un solo punto (para lo cual, habría que limitar al operador que se presentará a continuación).

La recombinación ocurrirá con una probabilidad p_c, con $0 \leq p_c \leq 1$. Por lo tanto, algunas cadenas pasarán a formar parte de la siguiente generación tal cual estaban en la generación anterior. A la no recombinación también se le llama clonación y su contribución a la siguiente generación será:

$$[P_I(t+1)]_{clon} = (1 - p_c) P_I^{'}(t) \tag{4.2}$$

Para describir los casos de sí recombinación debemos definir antes el concepto de *máscara*. En una recombinación de dos "padres"[1] una máscara m indica de qué manera se va a llevar a cabo la recombinación, es decir, qué bit se tomará de qué padre para "formar" al descendiente. La máscara será una cadena binaria de igual longitud a la cadena I (en caso de cadenas de longitud fija). Un 0 en cualquier posición de la máscara implicará que en esa posición, el descendiente tendrá el mismo bit que el primer padre, un 1, implicará que tendrá el mismo bit que el segundo padre en esa posición. Por ejemplo, la máscara 001 quiere decir que se tomarán los primeros dos bits del primer padre y el tercer bit del segundo padre, mientras que la máscara 101 implica que la primera y la tercera posición vendrán del segundo padre y la segunda posición del primer padre. En este trabajo, únicamente se analizará el caso de entrecruzamiento en un solo punto ("one-point crossover"), con lo cual, la máscara 101 y cualquier otra que intercale 0's y 1's quedará excluida. Para una cadena de tres bits, tenemos entonces sólo 4 máscaras posibles para cruza en un sólo punto: $001, 011, 100$ y

[1] Modelos aún más generales incorporan recombinación entre más de dos padres.

110 (formalmente también deben incluirse la 000 y la 111 aunque en realidad representan casos de no recombinación).

Ahora bien, como debemos calcular $P_I(t+1)$, debemos limitar de alguna manera todas las recombinaciones posibles a aquéllas que dan como resultado el esquema I. Supongamos que tenemos dos padres representados por los esquemas K y L, y queremos conformar mediante una recombinación entre ellos al esquema I. $\lambda_I^{KL}(m)$ representa la probabilidad condicional de que los padres K y L se recombinen con la máscara m y den lugar a un hijo I.

De esta manera, $\lambda_I^{KL}(m) = 1$ si y sólo si los padres K y L forman la cadena I recombinándose con la máscara m y $\lambda_I^{KL}(m) = 0$ si no la forman. Por ejemplo, $\lambda_{010}^{011,110}(001)$ indica que se tomen los primeros dos bits del primer padre (01) y el tercer bit del segundo padre (0) y se verifique si con ellos se construye la cadena 010. Como sí se forma, $\lambda_{010}^{011,110}(001) = 1$. Análogamente, $\lambda_{010}^{011,110}(011) = 1$, pero $\lambda_{010}^{011,110}(100) = 0$ y $\lambda_{010}^{011,110}(110) = 0$.

Por lo tanto, en un evento de recombinación, la probabilidad de generar al esquema I estará dado por:

$$\sum_{K,L} \lambda_I^{KL}(m) P_K^{'}(t) P_L^{'}(t) \tag{4.3}$$

Sin embargo, la ecuación 4.3 debe limitarse pues en cada evento de recombinación se elegirá exclusivamente una máscara, con lo cual la probabilidad de generar al esquema I deberá multiplicarse por la probabilidad condicional de elegir a la máscara m a la hora de recombinar y sumar sobre todas las m (ver sección 5.2), resultando así:

$$\sum_{m} p_c(m,t) \sum_{K,L} \lambda_I^{KL}(m) P_K^{'}(t) P_L^{'}(t) \tag{4.4}$$

Finalmente, recordemos que la recombinación no ocurre siempre, con lo cual debemos multiplicar 4.4 por p_c obteniendo así:

$$[P_I(t+1)]_{rec} = p_c \sum_{m} p_c(m,t) \sum_{K,L} \lambda_I^{KL}(m) P_K^{'}(t) P_L^{'}(t) \tag{4.5}$$

4.1.3. Mutación

La mutación consiste en la "transformación" de un esquema en otro.

En nuestro caso, consideraremos el caso más sencillo donde exclusivamente se permiten mutaciones puntuales, es decir, la modificación de un bit (de 0 a 1 o viceversa). La mutación de un bit es independiente de las mutaciones de bits adyacentes. Además, todas las posiciones tienen la misma probabilidad de mutación p_m y todos los cambios (en este caso sólo son 2, pero podría extenderse a más en caso de codificación no binaria) son igualmente probables.

Dado que sólo consideraremos mutaciones puntuales, la transformación del esquema J al esquema I dependerá del número de diferencias puntuales que

existan entre las cadenas. A esta diferencia se le conoco como la distancia Hamming (ver sección 3.2.3) y la denotaremos como $d_H(J,I)$. Si la longitud de las cadenas es l, el número de posiciones concordantes entre las cadenas J e I será $l - d_H(J,I)$.

Siguiendo el argumento presentado en [36]: dadas las condiciones de probabilidades independientes e iguales entre cada posición, la probabilidad de cambiar todas las posiciones diferentes entre las cadenas J e I será $(p_m)^{d_H(J,I)}$, mientras que la probabilidad de no cambiar las posiciones concordantes será $(1 - p_m)^{l-d_H(J,I)}$. Siendo así, la probabilidad de que el esquema J se transforme en el esquema I será:

$$p(J \to I) = (p_m)^{d_H(J,I)} \times (1 - p_m)^{l-d_H(J,I)} \tag{4.6}$$

Podemos entonces de definir la matriz de mutación M_I^J de dimensión $2^l \times 2^l$ con las probabilidades de transformación de cualquier esquema a otro, dadas por la ecuación 4.6. Así:

$$M_I^J = (p_m)^{d_H(J,I)}(1 - p_m)^{l-d_H(J,I)} \tag{4.7}$$

Dado que la probabilidad de que una cadena cualquiera se transforme en cualquiera de la otras (incluída sí misma) debe ser 1, tendremos que $\sum_I M_I^J = 1$ y $\sum_J M_I^J = 1$. Además la matriz será simétrica pues la transformación en ambos sentidos es igualmente probable.

4.1.4. Modelo completo

Ya habiendo descrito a los tres operadores fundamentales, podemos ahora combinarlos para calcular el valor esperado de la frecuencia genotípica de I al tiempo $t+1$. Si sumamos los términos de clonación (4.2) y recombinación (4.5), ambos aplicados a la población ya seleccionada, los multiplicamos por la matriz de mutación M_I^J y sumamos sobre J, obtenemos:

$$\langle P_I(t+1)\rangle = \sum_J M_I^J \left((1 - p_c)P'_J(t) + p_c \sum_m p_c(m,t) \sum_{K,L} \lambda_J^{KL}(m) P'_K(t) P'_L(t) \right) \tag{4.8}$$

La ecuación 4.8 toma en cuenta todas las posibilidades para formar un individuo con el genotipo I en el tiempo $t+1$ a partir de la población seleccionada al tiempo t, mediante clonación o recombinación y mutación. Nótese que también puede escribirse de la siguiente forma:

$$\langle P_I(t+1)\rangle = \sum_J M_I^J \left(P'_J(t) - p_c \sum_m p_c(m,t) \Delta_J(m,t) \right) \tag{4.9}$$

donde

$$\begin{aligned}\Delta_J(m,t) &= P'_J(t) - \sum_{K,L} \lambda_J^{KL}(m) P_K^{'}(t) P_L^{'}(t) \\ &= P_J^{'}(t) - P_{J_m}^{'}(t) P_{J_{\overline{m}}}^{'}(t)\end{aligned} \tag{4.10}$$

recuperando así, la notación que incorpora al desequilibrio de ligamiento tal como se describió en la sección 3.4.3.

Capítulo 5

La evolución de la recombinación

En la genética de poblaciones tradicional, la recombinación se da considerando a los genes como entes indivisibles, con tan sólo la variación correspondiente a sus distintos alelos. Los genes se intercambian como si fueran "cuentas en un collar", es decir, suponiendo que la recombinación se da únicamente en las fronteras de los genes (recombinación inter-génica).

De hecho, cuando se desarrollaron los primeros modelos de la recombinación aún no existía el concepto de "adentro de un gen" como tal. De cierta manera, la evidencia experimental que se tenía entonces apuntaba en la dirección de que para fines de la recombinación, la estructura interna de los genes era irrelevante. De hecho, ésto sigue siendo cierto hasta cierto punto, pues la mayor parte de la recombinación homóloga que se observa hoy en día, en efecto se da entre los genes [27].

Hoy sabemos, sin embargo, que los genes sí tienen estructura interna y que por lo tanto, desde un punto de vista microscópico la recombinación podría darse perfectamente "adentro" de los genes (recombinación intra-génica). Aunque en menor medida, la recombinación intra-génica sí existe y es factible pensar que la distribución de uso de puntos de cruza no siempre ha sido igual. En cualquier caso, es fundamental entender cómo y por qué se desarrollaron los mecanismos moleculares que favorecen la recombinación inter-génica.

Creemos que el tipo de recombinación homóloga que observamos hoy es el resultado de la evolución de la recombinación. Los mecanismos moleculares que observamos podrían verse como una consecuencia de la evolución de la recombinación y no la causa de su estado actual. Creemos que la recombinación pudo haberse adaptado para actuar preferentemente en algunas zonas del genoma y que estas zonas se fueron modificando a la par de la conformación de la estructura del genoma. Qué caracteriza a estas zonas y cuál podría ser la forma de un operador de recombinación que permitiera esta evolución serán las preguntas guía de lo que viene a continuación.

5.1. Exploración vs. explotación

Los términos "exploración" y "explotación" se utilizan comúnmente en la literatura de cómputo evolutivo, pero por lo general no se brinda una definición explícita de estos conceptos sino que se utilizan basándose en el significado intuitivo que les damos [7]. De hecho son utilizados regularmente en estudios relativos a procesos adaptativos incluso en áreas como el aprendizaje en instituciones. Vale la pena citar a uno de los máximos estudiosos en este ámbito, James G. March [20], quien dijo al respecto:

> "La exploración incluye aquello capturado en términos como búsqueda, variación, toma de riesgos, experimentación, juego, flexibilidad, descubrimiento, innovación. La explotación está capturada en refinamiento, elección, producción, eficiencia, selección, implementación, ejecución. Los sistemas adaptativos que sólo llevan a cabo la exploración a reserva de la explotación probablemente sufrirán los costos de experimentación sin tener acceso a sus beneficios. Exhibirán demasiadas ideas nuevas subdesarrolladas y muy poca competencia distintiva. Por otro lado, los sistemas que llevan a cabo explotación de forma exclusiva se encontrarán atrapados en puntos de equilibrio estables subóptimos. Como resultado de ello, mantener un balance apropiado entre la exploración y la explotación es un factor primario en la supervivencia y prosperidad del sistema."[1]

A pesar de que la cita anterior provenga de un contexto institucional, prácticamente todos los términos utilizados tienen sentido en el contexto evolutivo y sin duda, complementan nuestra intuición de los significados de exploración y explotación. Según Eiben y Schippers [7] no existe un consenso entre autores en torno a la exploración y explotación en los algoritmos evolutivos (que incluyen a los algoritmos genéticos y el cómputo evolutivo). La mayoría coincide en que la clave para la búsqueda eficiente de un óptimo es un equilibrio adecuado entre la exploración y la explotación. La convergencia prematura (a máximos locales) debería evitarse hasta no haber cubierto (explorado) la mayor proporción posible del espacio de búsqueda[2]. Siendo así, en la primera fase de la búsqueda, la exploración debe favorecerse. Conforme se vaya avanzando (en el tiempo) debe comenzarse a intentar sacar el máximo provecho de las mejores soluciones encontradas hasta el momento, es decir, debe favorecerse la explotación.

Vale la pena mencionar una diferencia sustancial entre los objetivos de la evolución en los algoritmos genéticos y en la evolución natural, pues el significado de "exploración" depende fuertemente de ellos. En los algoritmos genéticos, el objetivo es encontrar la solución óptima. En ese sentido, la explotación a la que se refieren en muchos casos se trata más bien de una exploración local, pues una vez encontrado una solución adecuada, su explotación consiste en encontrar el óptimo. Por otro lado, en la evolución natural el objetivo es aumentar

[1] Traducción del autor.

[2] En general, en la realidad, los espacios de búsqueda son tan grandes que se acaba explorando una región muy reducida del espacio de búsqueda total.

la adecuación promedio de la población. Para hacerlo, evidentemente también conviene encontrar una solución adecuada, pero la explotación se refiere no tanto a la exploración local (que se incluye dentro del término exploración) sino a la dispersión de la solución adecuada en la población de tal forma que la adecuación promedio de la población aumente.

Según Sa *et al.* [39], el papel de los operadores genéticos es asegurar que existe suficiente presión para obtener mejores soluciones a partir de las buenas soluciones ya encontradas (explotación) y cubrir el espacio de soluciones lo suficiente para maximizar la probabilidad de encontrar el óptimo global (exploración).

Algunos autores consideran que la explotación es el uso de información existente. Otros restringen el término explotación al uso de información buena o, de forma más refinada, al "buen uso de información" [12].

En cuanto a la exploración, las opiniones son aún más diversas. Algunos consideran que los operadores de mutación y recombinación son de naturaleza exclusivamente exploradora. Nuestra visión es mucho más cercana a la expuesta por Spears [34] quien considera que la mutación sirve para generar diversidad aleatoria en la población, mientras que la recombinación sirve como un acelerador que promueve comportamiento emergente a partir de sus componentes. En nuestra opinión, la recombinación no sólo es un operador de exploración sino que también es fundamental en la explotación (en evolución natural).

Si pensamos en la posibilidad de recombinar dos módulos óptimos estamos evidentemente ante un caso en que la recombinación contribuye a la explotación de dichos módulos. Un aumento en la tasa de recombinación podría presentarse como una solución, sin embargo, el aumento en la recombinación no siempre resulta favorable para la población pues siempre irá acompañado por un aumento en la disrupción (de módulos ya formados) [6]. Por otro lado, la disrupción no siempre es desfavorable. Si pensamos en una población convergente (donde la variación entre los individuos es cada vez menor), la recombinación será cada vez menos disruptiva pues padres similares tendrán descendencia similar a ellos. Paralelamente a esta transición, la naturaleza de la recombinación pasará de ser exploradora a ser explotadora [7]. La mutación, en cambio, siempre mantendrá su naturaleza exploradora y disruptiva, independientemente del individuo sobre el cual opere y del estado de la población.

5.2. Evolución en la probabilidad diferencial de recombinación

Un término de la ecuación 4.8 que vale la pena analizar con detenimiento es la probabilidad diferencial de recombinación representada por $p_c(m,t)$. La dependencia temporal de esta probabilidad es fundamental para la hipótesis central de este trabajo y la podemos definir de distintas formas. Podemos establecer que la probabilidad de entrecruzamiento en un punto determinado de la cadena (representado por la máscara m) dependa del efecto que el uso de dicha

m=0001			
F0_A	F0_B	F1	adec.
0***	*000	0000	0
0***	*001	0001	0
0***	*010	0010	0
0***	*011	0011	1
0***	*100	0100	0
0***	*101	0101	0
0***	*110	0110	0
0***	*111	0111	1
1***	*000	1000	0
1***	*001	1001	0
1***	*010	1010	0
1***	*011	1011	1
1***	*100	1100	1
1***	*101	1101	1
1***	*110	1110	1
1***	*111	1111	2
total			8

m=0011			
F0_A	F0_B	F1	adec.
00**	**00	0000	0
00**	**01	0001	0
00**	**10	0010	0
00**	**11	0011	1
01**	**00	0100	0
01**	**01	0101	0
01**	**10	0110	0
01**	**11	0111	1
10**	**00	1000	0
10**	**01	1001	0
10**	**10	1010	0
10**	**11	1011	1
11**	**00	1100	1
11**	**01	1101	1
11**	**10	1110	1
11**	**11	1111	2
total			8

m=0111			
F0_A	F0_B	F1	adec.
0***	*000	0000	0
0***	*001	0001	0
0***	*010	0010	0
0***	*011	0011	1
0***	*100	0100	0
0***	*101	0101	0
0***	*110	0110	0
0***	*111	0111	1
1***	*000	1000	0
1***	*001	1001	0
1***	*010	1010	0
1***	*011	1011	1
1***	*100	1100	1
1***	*101	1101	1
1***	*110	1110	1
1***	*111	1111	2
total			8

Figura 5.1: Cada tabla muestra el resultado de someter una población infinita homogénea, con B/L=2/4, a recombinación con máscara m en un paisaje AP (con la aguja representada por el genotipo 11. Los individuos $F0A$ y $F0B$ se recombinan con máscara m para generar al descendiente $F1$. La notación de esquemas permite representar a todas las posibles recombinaciones de forma reducida. La adecuación promedio de la población es $8/16 = 0.5$ independientemente de la máscara utilizada, como es de esperarse.

máscara tenga sobre su descendencia. Entre más aptos sean los individuos de la población al tiempo $t+1$ resultado de una recombinación con la máscara m (de dos individuos de la población al tiempo t) mayor será su probabilidad de subsistir y por lo tanto, mayores serán los beneficios de recombinar con la misma máscara al tiempo $t+1$. Lo que ocurrirá en consecuencia es una heredabilidad del uso en los puntos de cruza a pesar de que ésta no se hereda como tal. El cambio en la conformación de la población (con mayor representatividad de individuos con módulos óptimos) provocará una evolución en la distribución de la probabilidad del uso de máscaras.

Esta dependencia la podemos representar mediante la siguiente ecuación:

$$p_c(m, t+1) = \frac{f(m, t+1)}{\overline{f(t+1)}} p_c(m, t) \tag{5.1}$$

donde $f(m, t+1)$ es la adecuación de los individuos resultado de recombinar a la población presente al tiempo t con la máscara m y $\overline{f(t+1)}$ la adecuación promedio de toda la descendencia, independientemente de su *máscara madre*[3].

Con fines ilustrativos, que de ninguna manera prentenden ser exhaustivos, pongamos un ejemplo. Supongamos una población infinita de individuos representados por cadenas de longitud $l = 4$ con $B/L = 2/4$ evolucionando en paisaje adaptativo AP (con la aguja representada por el genotipo 11).

Considerando únicamente el entrecruzamiento en un solo punto, tendremos, pues, cuatro máscaras posibles: $m = 0001, 0011, 0111, 1000$. Utilizando la notación de esquemas podemos calcular fácilmente la adecuación promedio de la población. Como lo muestra la tabla de la figura 5.1, la adecuación promedio de la población al tiempo $t = 1$ (que llamamos $F1$) es $8/16 = 0.5$ indepen-

[3]Llamaremos *máscara madre* a la máscara utilizada para dar origen a un individuo determinado.

dientemente de la máscara utilizada, como es de esperarse, dado que estamos empezando con una población homogénea.

		m=0001		m=0011		m=0111		m=1000		m=1100		m=1110	
F1_A	F1_B	F2	adec.	F2	adec.	F2	adec.	F2	adec.	F2	adec.	F2	adec.
0011	0011	0011	1	0011	1	0011	1	0011	1	0011	1	0011	1
	0111	0011	1	0011	1	0111	1	0011	1	0111	1	0111	1
	1011	0011	1	0011	1	0011	1	1011	1	1011	1	1011	1
	1100	0010	0	0000	0	0100	0	1011	1	1111	2	1101	1
	1101	0011	1	0001	0	0101	0	1011	1	1111	2	1101	1
	1110	0010	0	0010	0	0110	0	1011	1	1111	2	1111	2
	1111	0011	1	0011	1	0111	1	1011	1	1111	2	1111	2
0111	0011	0111	1	0111	1	0011	1	0111	1	0011	1	0011	1
	0111	0111	1	0111	1	0111	1	0111	1	0111	1	0111	1
	1011	0111	1	0111	1	0011	1	1111	2	1011	1	1011	1
	1100	0110	0	0100	0	0100	0	1111	2	1111	2	1101	1
	1101	0111	1	0101	0	0101	0	1111	2	1111	2	1101	1
	1110	0110	0	0110	0	0110	0	1111	2	1111	2	1111	2
	1111	0111	1	0111	1	0111	1	1111	2	1111	2	1111	2
1011	0011	1011	1	1011	1	1011	1	0011	1	0011	1	0011	1
	0111	1011	1	1011	1	1111	2	0011	1	0111	1	0111	1
	1011	1011	1	1011	1	1011	1	1011	1	1011	1	1011	1
	1100	1010	0	1000	0	1100	1	1011	1	1111	2	1101	1
	1101	1011	1	1001	0	1101	1	1011	1	1111	2	1101	1
	1110	1010	0	1010	0	1110	1	1011	1	1111	2	1111	2
	1111	1011	1	1011	1	1111	2	1011	1	1111	2	1111	2
1100	0011	1101	1	1111	2	1011	1	0100	0	0000	0	0010	0
	0111	1101	1	1111	2	1111	2	0100	0	0100	0	0110	0
	1011	1101	1	1111	2	1011	1	1100	1	1000	0	1010	0
	1100	1100	1	1100	1	1100	1	1100	1	1100	1	1100	1
	1101	1101	1	1101	1	1101	1	1100	1	1100	1	1100	1
	1110	1100	1	1110	1	1110	1	1100	1	1100	1	1110	1
	1111	1101	1	1111	2	1111	2	1100	1	1100	1	1110	1
1101	0011	1101	1	1111	2	1011	1	0101	0	0001	0	0011	1
	0111	1101	1	1111	2	1111	2	0101	0	0101	0	0111	1
	1011	1101	1	1111	2	1011	1	1101	1	1001	0	1011	1
	1100	1100	1	1100	1	1100	1	1101	1	1101	1	1101	1
	1101	1101	1	1101	1	1101	1	1101	1	1101	1	1101	1
	1110	1100	1	1110	1	1110	1	1101	1	1101	1	1111	2
	1111	1101	1	1111	2	1111	2	1101	1	1101	1	1111	2
1110	0011	1111	2	1111	2	1011	1	0110	0	0010	0	0010	0
	0111	1111	2	1111	2	1111	2	0110	0	0110	0	0110	0
	1011	1111	2	1111	2	1011	1	1110	1	1010	0	1010	0
	1100	1110	1	1100	1	1100	1	1110	1	1110	1	1100	1
	1101	1111	2	1101	1	1101	1	1110	1	1110	1	1100	1
	1110	1110	1	1110	1	1110	1	1110	1	1110	1	1110	1
	1111	1111	2	1111	2	1111	2	1110	1	1110	1	1110	1
1111	0011	1111	2	1111	2	1011	1	0111	1	0011	1	0011	1
	0111	1111	2	1111	2	1111	2	0111	1	0111	1	0111	1
	1011	1111	2	1111	2	1011	1	1111	2	1011	1	1011	1
	1100	1110	1	1100	1	1100	1	1111	2	1111	2	1101	1
	1101	1111	2	1101	1	1101	1	1111	2	1111	2	1101	1
	1110	1110	1	1110	1	1110	1	1111	2	1111	2	1111	2
	1111	1111	2	1111	2	1111	2	1111	2	1111	2	1111	2
total			53		56		53		53		56		53

Figura 5.2: Partiendo de la misma población (representada por los individuos con adecuación distinta de 0 de la $F1$ de la tabla anterior) se llevó a cabo un entrecruzamiento entre cada posible par de individuos con máscara m, generando así a los individuos de la siguiente generación ($F2$). La recombinación con máscara inter-bloque, es decir, $m = 0011$ y $m = 1100$, da como resultado una población con mayor adecuación que las máscaras de recombinación intra-bloque. Asociar $p_c(m)$ a la adecuación de la descendencia resultado de un entrecruzamiento con máscara m tendrá, en este caso representativo, el efecto de aumentar la probabilidad de recombinar precisamente en las fronteras entre los bloques (y por consiguiente, aumentar la adecuación promedio de la población, relativa al uso de otras máscaras).

Si suponemos que sobre la población $F1$ se hace una selección proporcional[4] y se escogen únicamente los individuos con adecuación distinta de 0, obtendremos una población conformada por los individuos con genotipo 0011, 0111, 1011, 1100, 1101, 1110 y 1111. Si llevamos a cabo recombinación aleatoria entre dichos

[4]Si se hiciera selección proporcional estricta, tendríamos 2 individuos con genotipo 1111 por cada uno de los demás. Los resultados en ese caso son aún más claros, pero por simplicidad no se muestran.

individuos, podemos calcular el valor esperado de la adecuación en la generación siguiente dependiendo de la máscara utilizada, como se muestra en la tabla de la figura 5.2. Nótese que a pesar de que todos las poblaciones son iguales, la recombinación entre bloques, es decir, con máscaras $m = 0011$ y 1100 da una adecuación promedio mayor a la recombinación intrabloque. Si como indica la ecuación 5.1, la probabilidad de recombinación asociada a una máscara depende de la adecuación promedio de la descendencia producto de recombinar con dicha máscara, el resultado general y como se muestra en este ejemplo, será un aumento en la probabilidad de recombinar entre bloques. Dado un paisaje adaptativo fijo, las probabilidades de recombinación se adecuarán a dicho paisaje evitando, a medida que la población evoluciona, recombinar en puntos intrabloque (con efectos disruptivos) y privilegiando la recombinación interbloque (permitiendo la combinación de módulos óptimos en un mismo individuo).

A continuación se presentarán las ecuaciones de la recombinación adaptativa implementadas en nuestras simulaciones. La recombinación adaptativa implementada no dependerá explícitamente de la adecuación de los individuos resultado del entrecruzamiento en un determinado punto (como la representada por la ecuación 5.1), sino del número de bloques óptimos presentes en una determinada posición.

5.3. Recombinación de ruleta

En analogía a la selección de ruleta (*roulette wheel selection*), en la cual los individuos con mayor adecuación se seleccionan con mayor probabilidad para conformar la nueva población, la recombinación de ruleta elige puntos de cruza con mayor probabilidad si éstos tienen mayor adecuación recombinativa. Como se ha expuesto a lo largo de este trabajo, consideramos que una selección preferencial de puntos de cruza exploradores debería ser favorable en etapas tempranas de la evolución de la población, mientras que la selección de puntos de cruza explotadores debería ser beneficiosa una vez que los bloques óptimos se han encontrado.

En este sentido, la recombinación de ruleta es una recombinación adaptativa, pues elige diferentes puntos de cruza dependiendo del estado actual de la población. Para llevarla a cabo, se le asigna una adecuación a cada punto de cruza y se construye una ruleta con "áreas" proporcionales a las correspondientes adecuaciones recombinativas. Computacionalmente, el intervalo [0,1] se divide en fracciones (no sobrelapadas) proporcionales a las "áreas" de la ruleta. Para elegir el punto de cruza, se elige un número entre 0 y 1 y el intervalo en el que caiga indicará el punto en el cual se llevará a cabo el entrecruzamiento.

Consideraremos dos tipos diferentes de puntos de cruza: los inter-bloque y los intra-bloque. La adecuación recombinativa de los puntos de cruza inter-bloque se calculará a partir del número de bloques óptimos encontrados. Entre mayor sea el número de bloques óptimos encontrados, mayor será la probabilidad de recombinar en puntos inter-bloque y menor la probabilidad de recombinar en puntos intra-bloque. De esta forma pretendemos forzar el uso diferencial de

puntos de cruza y pasar de una recombinación predominantemente exploradora a una recombinación explotadora. Como se mencionó en la sección 5.1, conforme la población se va homogeneizando, el paso de recombinación exploradora a recombinación explotadora se da de forma natural, pues cada vez se pierde más el poder disruptivo de la recombinación. Con la recombinación de ruleta estamos forzando esta transición.

En este trabajo se utilizarán dos algoritmos de recombinación de ruleta diferentes: *Rul0* y *RulN*. La diferencia entre ambos algoritmos es pequeña pero significativa en los resultados. A continuación se presentarán las ecuaciones que los describen.

5.3.1. *Rul0*

La asignación de adecuaciones recombinativas a cada punto de cruza se lleva a cabo de la siguiente manera. En cada generación, se calcula la fracción de individuos F_{Inter_b} que tienen la adecuación máxima posible en el bloque número b.

En *Rul0* la adecuación recombinativa F_{Inter_b} asociada al punto inter-bloque entre los bloques b y $b+1$ es:

$$F_{Inter_b} = \sum_{i=1}^{N} opt_{i_b} \tag{5.2}$$

donde la suma sobre i es sobre todos los individuos de la población, $Inter_b$ representa la frontera del bloque b, con $b = 0, 1, \ldots, L/B$ (L es la longitud de la cadena y B es la longitud de los bloques) y

$$opt_{i_b} = \begin{Bmatrix} 1 & \text{si el individuo } i \text{ contiene al bloque óptimo } b \\ 0 & \text{de lo contrario} \end{Bmatrix}$$

A todos los puntos intra-bloque dentro del bloque b se les asigna la adecuación recombinativa $N - F_{Inter_b}$.

$$F_{Intra_{k_b}} = N - F_{Inter_b} = N - \sum_{i=1}^{N} opt_{i_b} \tag{5.3}$$

donde $k = 0, 1, \ldots, B$, corresponde a las posiciones dentro del bloque b.

El mismo procedimiento se lleva a cabo para cada bloque, con la excepción de que a la frontera del último bloque no se le asigna su correspondiente adecuación, pues un entrecruzamiento en este punto equivaldría a no recombinar. Las adecuaciones recombinativas se normalizan de tal manera que sumen 1 y se asignan de manera cumulativa a cada punto de cruza, conformándose así la ruleta. Estas adecuaciones se calculan una vez que los operadores de selección, recombinación y mutación se hayan aplicado.

5.3.2. *RulN*

En el algoritmo *RulN*, la adecuación de los puntos inter-bloque es:

$$F_{Inter_b} = N + \sum_{i=1}^{N} opt_{i_b} \tag{5.4}$$

donde $b = 0, 1, \ldots, L/B$ y N es el tamaño de la población. Además, a todos los puntos intra-bloque dentro del bloque b se les asigna la adecuación recombinativa $N - F_{Inter_b}$ al igual que en *Rul0*:

$$F_{Intra_{k_b}} = N - \sum_{i=1}^{N} opt_{i_b} \tag{5.5}$$

donde $k = 0, 1, \ldots, B$, corresponde a las posiciones dentro del bloque b.

La diferencia entre los *Rul0* y *RulN* radica únicamente en que en *Rul0* la adecuación de los puntos inter-bloque comienza siendo 0, mientras que en el *Rul1* comienza siendo N (de ahí la elección de nombres). Las repercusiones de esta diferencia se comentarán en el capítulo 6.

Capítulo 6

Simulaciones

En este capítulo se presentarán los resultados de algunas de las simulaciones ejecutadas con el propósito de mostrar que una recombinación adaptativa que favorece la formación y preservación de módulos, le confiere a las poblaciones una ventaja adaptativa respecto aquéllas cuyo modelo de recombinación es aleatorio y no relacionado con la estructura modular del individuo.

La intención original de este capítulo era llevar a cabo esta misión mediante la comparación de diferentes modelos de recombinación en distintos paisajes de adecuación con variación de parámetros. Sin embargo, el enfoque cambió durante el desarrollo del trabajo pues las simulaciones en paisajes DAP resultaron mucho más complejas e interesantas de lo esperado y por lo tanto se profundizó más en algunos detalles que en un principio ni siquiera se habían considerado. Una de las repercusiones de esto fue no haber explorado la evolución en paisajes de adecuación con epistasis aditiva intra-bloque, como el paisaje CU, o no haber podido explorar detenidamente el papel de la mutación en las simulaciones, por lo cual, los resultados mostrados en este capítulo son todos en ausencia de mutación.

El capítulo está dividido en tres secciones. En la sección 6.1 se mencionan algunos parámetros relevantes de las simulaciones, mientras que en las secciones 6.2 y 6.3 se muestran los resultados de la simulaciones en paisajes AP y DAP, respectivamente. En la sección 6.3.1 se analiza la *coexistencia de agujas*, que se puede relacionar con la existencia de más de un alelo segregando en un mismo locus, y en la sección 6.4, se describe la evolución en el uso de puntos de cruza.

6.1. Parámetros

En ausencia de mutación, uno de los aspectos que más influenciará el desempeño de nuestros algoritmos de ruleta es la representatividad de bloques de alta adecuación en la población inicial. Tener un número inical de bloques óptimos beneficiará al ritmo evolutivo (provocará un aumento en la adecuación promedio de la población más rápido) de los algoritmos de ruleta en comparación con el

algoritmo de recombinación aleatoria. Esto se debe a que mientras el algoritmo de recombinación aleatoria explotará estos bloques únicamente a través de la selección, los algoritmos de ruleta privilegiarán la recombinación inter-bloque, favoreciendo también con este mecansimo la explotación de los bloques óptimos y acelerando el ritmo evolutivo.

En ese sentido hay dos parámetros importantes en nuestras simulaciones. Éstos son el tamaño poblacional N y la longitud de los bloques B. Entre mayor sea el tamaño poblacional, mayor será la probabilidad de que bloques óptimos estén presentes en la población inicial. Por otro lado, entre menor sea el tamaño del bloque, menos esquemas posibles existirán para cada bloque y por lo tanto, mayor será la probabilidad de que uno elegido al azar sea óptimo. Dado que el número de esquemas posibles de un bloque de tamaño B es 2^B, la relación entre N y 2^B será el parámetro relevante.

En un principio se había decidido mantener fijo el número de bloques por cadena para evitar tener otro parámetro libre. Se eligieron las poblaciones cuya razón entre la longitud de total de la cadena L y la longitud de cada bloque B fuera igual a 4, en particular, $B/L = 4/16$ y $B/L = 8/32$. Sin embargo, algunas observaciones nos llevaron a experimentar con diferentes números de bloques por cadena (sección 6.4).

No está de más recordar que nuestras simulaciones incorporan dos eventos aleatorios importantes: la generación aleatoria de la población inicial, por un lado, y por otro, la deriva génica, es decir, la pérdida aleatoria de individuos. La deriva génica es una consecuencia de nuestro algoritmo de selección que elige parejas de individuos aleatoriamente de la población y elige al de mayor adecuación. Un individuo de alta adecuación, sin embargo, podría nunca ser elegido para *competir* en cuyo caso quedaría eliminado de la población por deriva génica.

6.2. Aguja en un pajar

Como se comentó en la sección 3.2.2, en el paisaje de adecuación de Aguja en un pajar (AP), todos los genotipos (o cadenas) tienen igual adecuación excepto uno, que tiene una adecuación mayor. Nosotros hemos elegido el paisaje AP donde la aguja tiene una adecuación de 2, mientras que el resto tiene una adecuación de 1.

En la figura 6.1 se muestran los resultados que obtuvimos al comparar los tres algoritmos de recombinación diferentes: recombinación aleatoria (*Al*), recombinación de ruleta *Rul0* (definido en la sección 5.3.1) y recombinación de ruleta *RulN* (definido en la sección 5.3.2).

Dado que las diferencias entre las curvas de la figura 6.1 son difíciles de percibir y que a lo largo de este trabajo se querrán mostar reiteradamente las diferencias en el desempeño de varios algoritmos, optaremos a partir de ahora por graficar la diferencia (la resta) entre las adecuaciones promedio de las poblacaciones entre un algoritmo y otro, y no la adecuación como tal. De esta manera será mucho más fácil analizar las curvas pues que sean positivas o negativas en

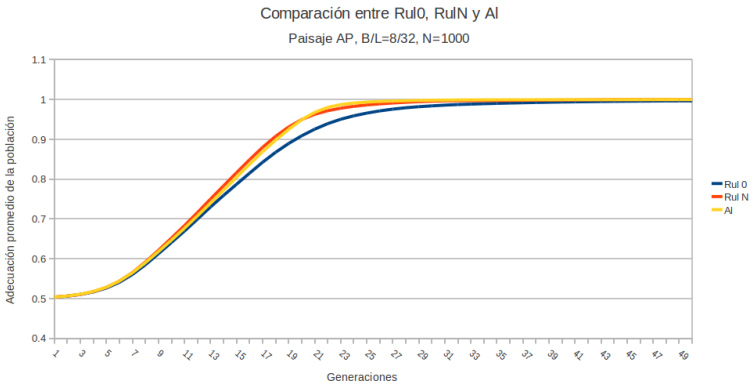

Figura 6.1: Se simularon 500 poblaciones de 1000 individuos cada una ($B/L = 8/32$) (con selección de torneo y sin mutación) con recombinación *Rul0*, *RulN* y *Al* en un paisaje AP. La adecuación promedio de cada población está normalizada de tal manera que la adecuación máxima de la población sea 1. La similitud entre las curvas *RulN* y *Al* evidencia que las diferencias entre estos 2 algoritmos son pequeñas en este paisaje, mientras que el algoritmo *Rul0* tiene un desempeño claramente inferior.

uno u otro intervalo indicarán un mejor desempeño de una algoritmo frente al otro en dicho intervalo.

En la figura 6.2 se muestra una comparación entre los aloritmos *Rul0* y *RulN* para tamaños de población de $N = 500$, $N = 1000$ y $N = 2000$. Como se comentó en la sección 6.1 el tamaño de población es un parámetro importante para el desempaño del algoritmo. Un análisis detallado y cuidadoso de estas gráficas nos permite conocer el comportamiento de nuestros algoritmos a un nivel difícil de acceder nada más a partir del conocimiento de las ecuaciones que describen su evolución.

Por ejemplo, en el paisaje AP, independientemente del tamaño poblacional, el algoritmo *Rul0* es claramente peor que el algoritmo *Al*, que recombina en puntos aleatoriamente elegidos en cada evento de recombinación. A pesar de que nuestra intuición nos indica que recombinar preferencialmente en puntos intra-bloque al inicio y luego recombinar en puntos inter-bloque debería incrementar el ritmo evolutivo de la población, esto no ocurre. La transición en uso de puntos de cruza de intra-bloque a inter-bloque en efecto se lleva a cabo (como muestra la gráfica **b** de la figura 6.2), sin embargo, esto no representa ventaja alguna respecto al algoritmo que recombina de forma aleatoria. Este hecho nos indica que el carácter explotador de la recombinación es sumamente importante en todo momento.

La gráfica **b** de la figura 6.2 nos muestra la evolución en uso de puntos de cruza (como promedio de 500 realizaciones). Lo que estas curvas nos muestran de forma indirecta es la evolución en las adecuaciones recombinativas de cada algoritmo. Evidentemente, para *Al*, la fracción de uso de puntos inter-bloque e intra-bloque es constante e igual a 3/31 y 28/31 respectivamente (la probabilidad de recombinar es igual para todos los posibles puntos de entrecruzamiento).

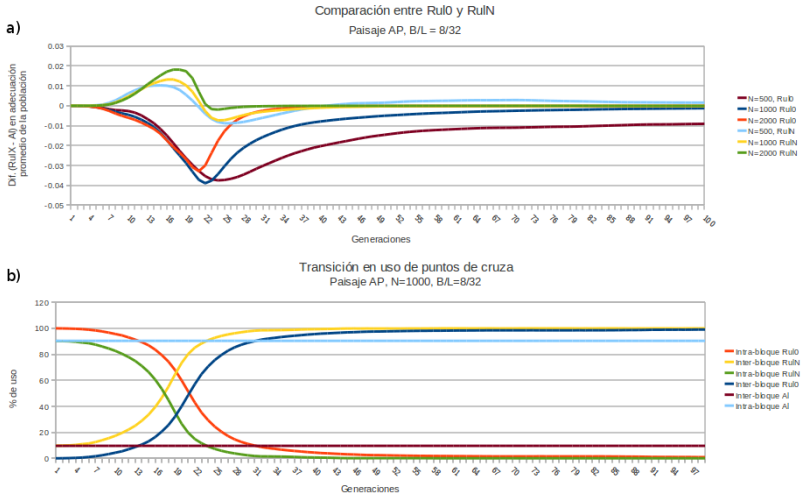

Figura 6.2: La gráfica **a** compara el resultado de los algoritmos *Rul0* y *RulN* respecto a *Al*. Se simularon 4000 poblaciones de 500, 1000 y 2000 individuos en un paisaje AP con B/L = 8/32, utilizando cada algoritmo (sin mutación). Para ambos algoritmos se muestra que una mayor representatividad de bloques óptimos en la población inicial conlleva que éstos se esparzan de forma más rápida en la población, incrementando así el desempeño de los algoritmos al aumentarse la población. Además, se muestra la clara superioridad del algoritmo *RulN* respecto a los otros dos, debido a la efectividad en su utilizacion de puntos de cruza. La gráfica **b** muestra la evolución en el uso de puntos de cruza con cada algoritmo, resultado del promedio de 500 simulaciones con $N = 1000$ y $B/L = 8/32$. La diferencia entre las curvas correspondientes a los algoritmos *Rul0* y *RulN* denota que aprovechar el carácter explotativo de la recombinación (entrecruzamiento inter-bloque) desde un principio tiene repercusiones importantes en el ritmo evolutivo de la población.

Para *Rul0* toda la recombinación comienza siendo intra-bloque, mientras que en *RulN*, la probabilidad de recombinar comienza siendo igual para todo punto como en *Al* y (conforme se van encontrando los bloques óptimos) se va pasando de recombinación predominantemente intra-bloque a recombinación inter-bloque. La comparación entre las gráficas **a** y **b** nos muestra que en efecto, el carácter explotativo de la recombinación juega un papel importante en todo momento y que no conviene dejar de recombinar en puntos inter-bloque en ningún momento. Sin embargo, la superioridad de *RulN* sobre *Al* muestra que nuestra intuición no estaba del todo incorrecta y que privilegiar el entrecruzamiento inter-bloque conforme se van encontrando bloques óptimos, aprovechando más el carácter explotativo de la recombinación y disminuyendo su poder explorativo es conveniente en términos evolutivos.

6.3. Doble aguja en un pajar

Sabemos que el desempeño de un algoritmo de búsqueda depende fuertemente del paisaje adaptativo en el cual evoluciona. En un piasaje DAP, donde existen dos esquemas óptimos (las agujas), la *ubicación* relativa de estos esquemas será determinante.

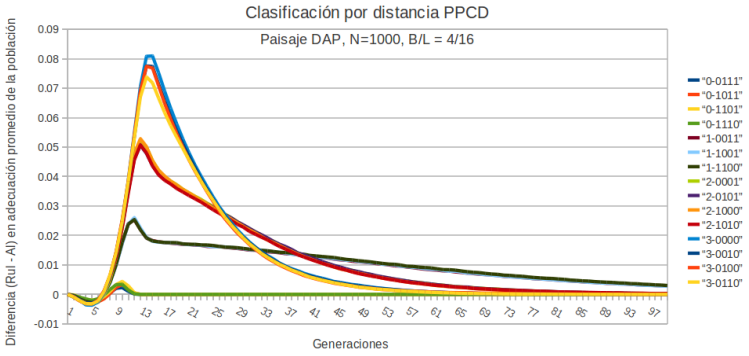

Figura 6.3: Se simularon 500 poblaciones de 1000 individuos cada una ($B/L = 4/16$), con selección de torneo y sin mutación, con recombinación de ruleta *Rul0* por un lado, y recombinación aleatoria por otro en diferentes paisajes DAP. La gráfica muestra la diferencia en adecuación promedio de las poblaciones simuladas con ambos métodos (*Rul0* menos *Al*) como función de la generación. Cada línea corresponde a un paisaje DAP diferente. La diferencia entre cada una es la ubicación de una de las agujas del paisaje (en todos, la otra aguja era la 1111). La leyenda indica la distancia PPCD entre ambas agujas del correspondiente paisaje y el esquema representativo de la aguja en cuestión ("distanciaPPCD-esquemaDeAguja"). La clara agrupación de las curvas dependiendo de su distancia PPCD indica que la clasificación de paisajes DAP con este criterio es muy adecuada en escenarios evolutivos sin mutación.

Como se describió en la sección 3.2.3 se pueden clasificar a todos los paisajes DAP a partir de las diferencias entre las cadenas que representan a cada "aguja". Las diferencias se pueden expresar mediante la distancia Hamming o mediante la distancia PPCD.

Un fenómeno que se esperaría observar en un paisaje DAP es la *coexistencia de agujas*. En nuestro caso, donde estamos tratando con paisaje modulares, una coexistencia de agujas sería importante cuando en diferentes individuos de la población, en un tiempo determinado, hubiera representatividad de ambas agujas en un bloque específico (el 1, 2, 3 ó 4).

Dada una coexistencia de agujas, una recombinación en algún punto inter-bloque podría provocar la disrupción de los esquemas representativos de cada aguja, disminuyendo la adecuación de los individuos involucrados en la recombinación. Siendo así, y como se comentó en la sección 3.2.3, entre menor sea el número de puntos de cruza que destrocen (PPCD) a los esquemas, menor será, digamos, el riesgo de la recombinación inter-bloque.

Si elegimos un paisaje DAP donde la recombinación entre agujas no cause daño alguno, es decir, donde no se deshagan las agujas, los eventos de recombinación disruptivos serán menos y por lo tanto, esperaríamos observar un aumento en el ritmo evolutivo. Contrariamente, la distancia Hamming entre agujas no debería ser determinante en un escenario sin mutación.

En la figura 6.3 se muestra el resultado de la evolución de 500 poblaciones de 1000 individuos ($B/L = 4/16$) en 15 paisajes DAP diferentes. Las curvas aparecen claramente agrupadas dependiendo de la distancia PPCD entre las agujas del paisaje DAP en el cual evolucionan. Esto demuestra que en un esce-

nario sin mutación, con coexistencia de agujas (ver sección 6.3.1), la distancia PPCD es un criterio mucho más adecuado de clasificación de paisajes DAP que la distancia Hamming.

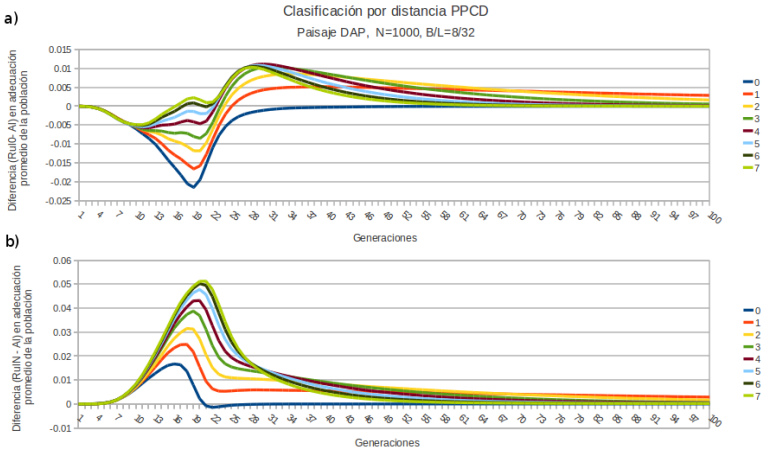

Figura 6.4: Se simularon 500 poblaciones de 1000 individuos cada una ($B/L = 8/32$), con selección de torneo y sin mutación, con recombinación de ruleta por un lado (*Rul0* para **a** y *RulN* para **b**) y recombinación aleatoria por otro en 255 diferentes paisajes DAP. Las gráficas muestra la diferencia en adecuación promedio de las poblaciones simuladas con ambos métodos (*Rul0-Al* y *RulN-Al*, para **a** y **b** respectivamente) como función de la generación. Cada curva representa el promedio de todas las simulaciones con la misma distancia PPCD entre agujas (de 0 a 7). La organización de las curvas correlacionada con su distancia PPCD es una muestra de que dicho criterio es adecuado para la clasificación. Por otro lado, se puede corroborar con estas gráficas el mejor desempeño del algoritmo *RulN* respecto a *Rul0* y *Al*.

A raíz de estos resultados, se llevó a cabo la misma prueba para poblaciones con $B/L = 8/32$ con distintos tamaños poblacionales. Se evolucionaron poblaciones en 255 paisajes DAP diferentes, correspondientes a todas las agujas posibles (el esquema 11111111 se eligió como la segunda aguja en todos los casos) con la misma población inicial y se promediaron los resultados de 1000 realizaciones tomando como criterio la distancia PPCD entre agujas. En la figura 6.4 se muestran los resultados de la simulación para $N = 1000$ (con $N = 500$ y $N = 2000$ se obtienen resultados cualitativamente iguales). La organización de las curvas dependiendo de la distancia PPCD muestra que en este caso también se trata de un buen criterio de clasificación.

6.3.1. Coexistencia de agujas

El hecho de que la clasificación por la distancia PPCD entre agujas sea adecuada, implica indirectamente que la coexistencia de agujas debe ser un fenómeno determinante en nuestras simulaciones. Remitiéndonos de nuevo a las figuras 6.3 y 6.4, cabe resaltar que la diferencia en el desempeño del algoritmo *Rul0* frente a *Al* es mayor cuando la distancia PPCD es mayor. Tomando en cuenta que entre mayor sea la distancia PPCD entre agujas, más *peligrosa* será la

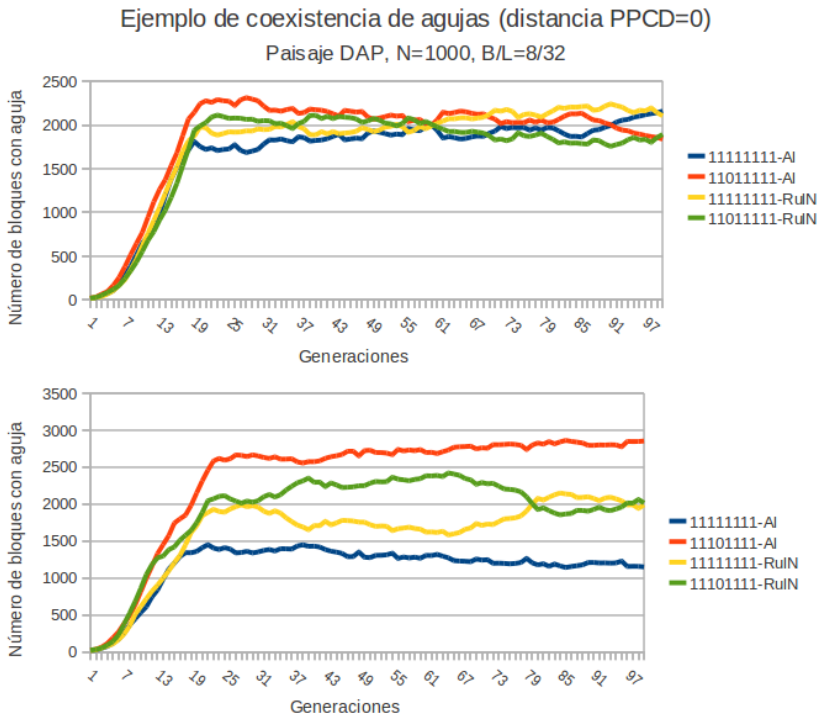

Figura 6.5: Ambas gráficas representan casos típicos de coexistencia de agujas tanto para *Al* como para *RulN*. El eje vertical cuenta el número de bloques óptimos presentes en la población. Como $N = 1000$ y el número de bloques es 4, cuando se alcanza la adecuación máxima, el número total de agujas debe ser 4000. Por lo tanto, la suma de la línea azul con la línea roja, por un lado, y la suma de la verde con la amarilla por otro, deben sumar 4000 a partir de cierto número de generaciones transcurridas (en este caso, alrededor de 20). La coexistencia de agujas en por lo menos uno de los bloques se deduce de esta imagen pues las curvas oscilan. Si no hubiera coexistencia, las curvas se mantendrían constantes e iguales a algún múltiplo exacto de 1000 a partir de cierto momento, indicando que todos los individuos contienen la misma aguja en cada posición (bloque). La coexistencia se permite en el algoritmo *Al* pues la distancia PPCD entre agujas es nula y por lo tanto, la recombinación intra-bloque no tiene efectos destructivos sobre las agujas.

recombinación intra-bloque y que para $N = 1000$, ambos algoritmos alcanzan la adecuación máxima en menos de 100 generaciones, es de suponerse que para distancias PPCD mayores a cero, la coexistencia de agujas en el algoritmo *Al* sea poca (y claramente menor conforme aumenta la distancia PPCD). Con recombinación de ruleta, en cambio, disminuye la probabilidad de recombinar intra-bloque conforme la adecuación aumenta (en términos generales) y por lo tanto, la coexistencia de agujas podría darse por tiempo indefinido[1].

La figura 6.5 muestra dos ejemplos característicos de coexistencia de agujas tanto con *RulN*[2] como con *Al* cuando la distancia PPCD entre agujas es

[1] En algún momento, debido a la deriva génica, una aguja específica acabaría por fijarse en cada bloque.

[2] *Rul0* da resultados cualitativamente iguales a *RulN*.

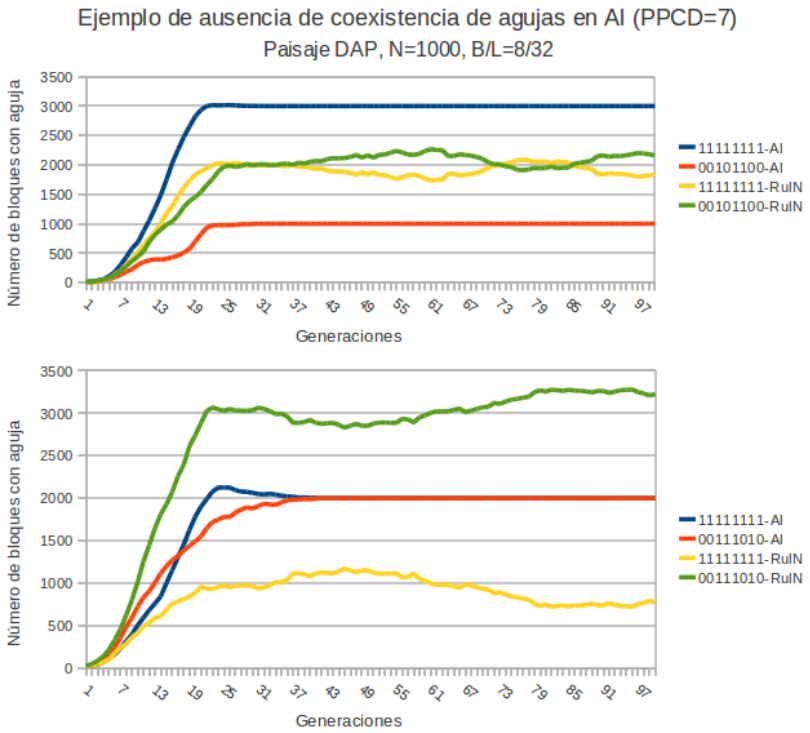

Figura 6.6: Estas gráficas representan casos típicos del caso contrario a la figura 6.5. Debido a que las distancia PPCD entre agujas es 7, cualquier recombinación intra-bloque las destruiría, y por lo tanto, la única manera para conseguir la máxima adecuación con el algoritmo *Al* es la "elección" de una de las dos agujas para cada bloque (indicada por el hecho de que las curvas azules y rojas se mantengan constantes e iguales a un múltiplo exacto de 1000 a partir de una generación determinada -que indica que se alcanzó la adecuación máxima). Para el algoritmo *RulN*, la coexistencia sí es posible, pues la recombinación intra-bloque disminuye progresivamente hasta volverse nula y posibilita el mantenimiento de ambas agujas en cualquiera de los bloques. Entendemos que un algoritmo como *RulN* permite la presencia simultánea dos (o más) alelos en la población a pesar de ser muy diferentes (relacionado con la distancia PPCD).

0. Cuando se alcanza la adecuación máxima, habrá una aguja en cada uno de los bloques en cada individuo de la población. Si la distancia PPCD es cero, la recombinación entre ambas agujas no será perjudicial y por lo tanto, se podrá alcanzar la adecuación máxima con coexistencia de agujas (recordemos que la coexistencia de agujas es la presencia simultánea dentro de la población de dos agujas diferentes en una posición específica).

Por otro lado, la figura 6.6 muestra dos ejemplos con distancia PPCD igual a 7, donde a diferencia de *RulN*, la coexistencia de agujas se pierde para *Al*.

6.4. Recombinación intrabloque vs. recombinación interbloque

El algoritmo *RulN* está construido de tal manera que la probabilidad de recombinar en puntos inter-bloque aumente a medida que se van encontrando módulos óptimos en los bloques ubicados justo antes del punto en cuestión. Si comenzamos con una distribución casi homogénea[3] de uso de puntos de cruza, podremos observar como ésta se irá modificando conforme pasen las generaciones. El la figura 6.7 se muestra el uso (real) de puntos de cruza como promedio de 500 realizaciones con poblaciones de 1000 individuos y $B/L = 8/32$, simuladas en paisajes DAP con 255 pares diferentes de agujas. A medida que van creciendo los "picos", indicando una mayor uso de puntos de cruza interbloque disminuye la recombinación en puntos intrabloque. Los "picos" crecerán hasta alcanzara probabilidad de 1/3, mientras que el resto dismuirá hasta hacerse 0.

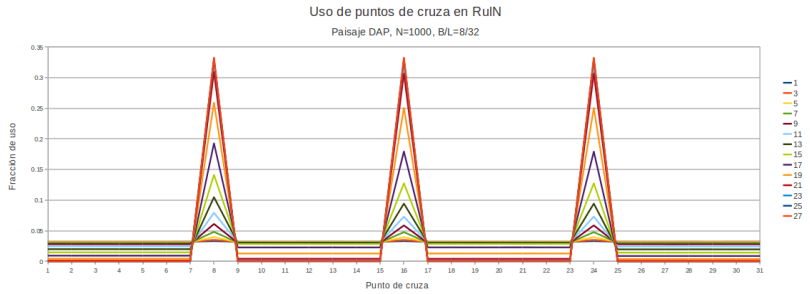

Figura 6.7: La gráfica muestra el cambio en la distribución de uso de puntos de cruza como promedio de 1000 realizaciones de poblaciones de 1000 individuos con B/L=8/32 evolucionando en 255 diferentes paisajes DAP. Cada curva corresponde a una generación y los picos correponden precisamente a los puntos de cruza inter-bloque. En la gráfica se aprecia como, a medida que pasan las generaciones, se va incrementando el entrecruzamiento en puntos interbloque y disminuyendo el entrecruzamiento intrabloque. El pico ubicado en el punto 8, que corresponde a recombinar entre el primer y segundo bloques es diferente a los picos en el 16 y 24 que son iguales entre ellos. De la misma forma, las curvas del punto 1 al punto 7 con iguales a las del punto 25 a 31, mientras que las curvas del 9 al 15 son iguales a las del punto 17 al 23, pero hay diferencias claras entre los bloques "fronterizos" y los bloques "centrales".

El algoritmo *RulN* establece iguales probabilidades de entrecruzamiento para todos los puntos (intrabloque) dentro de un mismo bloque, pero hay independencia entre bloque y bloque. Podría pensarse que la aparición de agujas en cada uno de los bloques será independiente de los demás. Sin embargo, esto no así. Como se muestra en la figura 6.7, la distribución de uso de puntos de cruza es igual entre los bloques 1 y 4[4], y entre los bloques 2 y 3, pero no entre

[3] Decimos "casi homogénea" porque en una población de $N = 1000$, una población aleatoria inicial tendrá un promedio cercano a 8 (exactamente $(2/(2^8)) * 1000 = 7.8125$) agujas en cada posición. Por lo tanto, la probabilidad de recombinar en puntos inter-bloque será ligeramente mayor al resto desde la primera generación.

[4] El pico al final del bloque 4 se calcula, y de ahí se obtienen las probabilidades de recombinar adentro del bloque 4, pero no se muestra dado que la probabilidad de recombinar en la frontera

los primeros y los segundos. En otros palabras, las agujas se encuentrán más rápidamente en los bloques "fronterizos" que en los bloques "centrales" y esto ocurre tanto con el algoritmo *RulN* como con el *Al*, como se muestra en la figura 6.8. Se corrieron varias simulaciones con los mismos parámetros y se observaron siempre los mismos resultados. Estaríamos pues frente a un un hecho que no habíamos predicho, a saber, que entre más lejano esté un bloque de las fronteras, mayor será la disrupción que sufrirá debido a la recombinación en otros bloques.

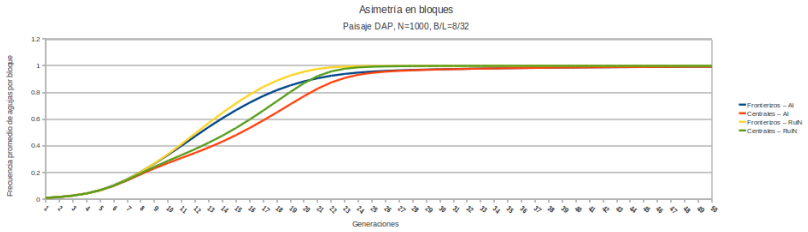

Figura 6.8: Esta gráfica muestra la fracción de agujas por bloque como función de la generación. Las curvas muestran la presencia de agujas en los bloques "fronterizos" (el 1 y el 4) y en los "centrales" (el 2 y el 3) bajo ambos algoritmos (*Al* y *RulN*). Las curvas son el resultado de promediar 1000 realizaciones de poblaciones de 1000 individuos con B/L=8/32, evolucionando en 255 diferentes paisajes DAP. A pesar de que al inicio las agujas se encuentran en igual proporción en todos los bloques, a partir de la generación 7, comienza a notarse una presencia mucho mayor de agujas en los bloques "fronterizos" en relación a los "centrales" en ambos algoritmos.

Para determinar si esta observación dependía de parámetros como el tamaño del bloque, se corrieron simulaciones con distintos tamaños de bloque y longitud de cadenas, también considerando números impares de bloques. En la figura 6.9 se muestra un ejemplo, que corresponde a poblaciones con $B/L = 5/40$, es decir, con 8 bloques en lugar de los 4 que se han simulado hasta ahora. En estas gráficas se muestra el mismo fenómeno, a saber, que la formación de agujas se da más rápidamente entre más cercano se encuentren los bloques a las fronteras. ¿Acaso no va esta observación en contra de lo esperado? La respuesta es que sí[5] Un paisaje de adecuación modular, como el que se está simulando aquí, donde la adecuación de un individuo es la suma de las adecuaciones de sus bloques, es por definición, un paisaje de adecuación aditivo y por lo tanto, sin epistasis. Adentro de los bloques, sin embargo, la epistasis es alta, recordemos, pues el paisaje de adecuación adentro de los bloques es un paisaje DAP. En principio, la epistasis intra-bloque no debaría afectar a la epistasis inter-bloque, con lo cual, esperaríamos que la razón de la disrupción de los bloques centrales se deba a la simulación del proceso evolutivo *per se*, quizás a una propiedad emergente del sistema. En [35] también se describe este efecto y se comenta que no se

de la cadena se iguala a 0. La semejanza entre los patrones de curvas intrabloque de los bloques 1 y 4, indica que el pico ubicado en la posición 8 es igual al pico teórico ubicado en la posición 32.

[5]Si se hace el cálculo del desequilibrio de ligamiento en un paisaje CU considerando a cada bloque un bit e ignorando la estructura interna de cada bloque (es decir, un paisaje sin epistasis entre las posiciones), el desequilibrio de ligamiento asociado al esquema 1*** es igual al asociado al esquema *1**. Nuestros resultados no concuerdan con dicho cálculo.

trata de un efecto directo de la recombinación (pues p_c no forma explícitamente parte de las ecuaciones de los bloques constructores de orden uno), sino uno indirecto, proveniente del valor real de la adecuación de cada esquema (en este caso, $f_{1***} > f_{*1**}$).

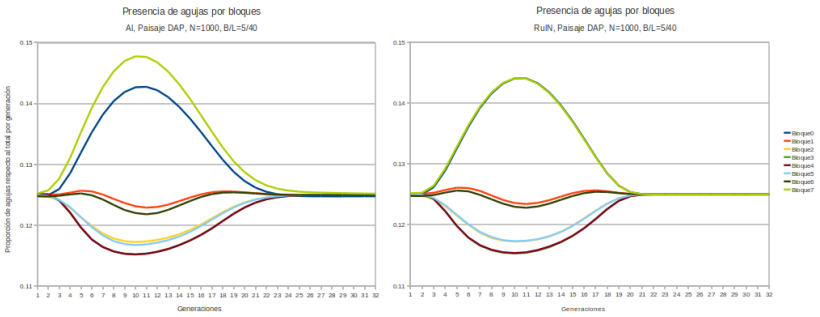

Figura 6.9: Estas gráficas muestran la proporción de agujas presentes en cada bloque en relación al total de agujas por generación. Ambas gráficas son el promedio de 2000 realizaciones de poblaciones de 1000 individuos con B/L=5/40, y cada realización llevada a cabo en 31 diferentes paisajes DAP. La gráfica de la izquierda es el resultado del algoritmo *Al* y la de la derecha, del algoritmo *RulN* (las poblaciones iniciales son iguales para ambos algoritmos). Nótese como a pesar de que todos los bloques empiezan con igual proporción de agujas, en los bloques "fronterizos" ésta aumenta mucho más rápidamente que en el resto. También es claro el orden en el cual se van formando agujas en el resto de los bloques, yendo digamos, de afuera hacia adentro, siendo los bloques "centrales" los más "lentos". Cabe mencionar que las regiones de las curvas con pendiente negativa, no implican que el número total de agujas en el bloque disminuya; el número total de agujas en cada bloque aumenta monotónicamente, de forma análoga a como se muestra en la gráfica 6.8. Llama la atención la diferencia entre las curvas de bloques simétricos en el algoritmo *Al*. En ambos casos, sin embargo, ocurre una disrupción de las regiones centrales, dando indicios de que a pesar de que los paisajes de adecuación son modulares (*i.e.* sin epistasis entre bloques), el comportamiento de los bloques sí depende de su ubicación dentro de la cadena indicando una epsitasis *de facto* entre bloques.

Observamos, pues, que la ubicación de los bloques en la cadena es sumamente relevante para determinar el ritmo en que se formarán los bloques óptimos. Cabe mencionar que el desequilibrio de ligamiento determinado por selección sólo nos muestra el efecto de la recombinación de una generación a la inmediatamente posterior y que considera poblaciones infinitas y completamente homogéneas. Por lo tanto, no nos aporta información para explicar el resultado de las simulaciones. La integración de las ecuaciones nos mostaría el equilibrio final de los bloques donde éstos se vuelven, de nuevo, indistinguibles, y es que, como se muestra en las figuras 6.7, 6.8 y 6.9, los efectos disruptivos de la recombinación se pueden observar durante un limitado número de generaciones. Un análisis profundo y detallado de esta fenomenología representa, sin duda, una interesante línea futura de investigación que por razones prácticas (y de salud mental) es imposible abarcar en este trabajo.

Capítulo 7

Conclusiones

En este planeta y en relación a la vida conocida existe sólo un tipo de evolución biológica, resultado del actuar simultáneo de un sinnúmero de fenómenos. Estos fenómenos se pueden clasificar de múltiples maneras y una sencilla es hacerlo en dos grupos: aquéllos relacionados con los cambios en las secuencias y aquéllos relacionados con la preservación de las secuencias en sí. En el primer grupo están incluidos todos los mecanismos moleculares responsables de la recombinación y la mutación, mientras que el segundo está representado por la selección natural, la deriva génica, los fenómenos demográficos, como cuellos de botella o el efecto fundador. La relación y *retroalimentación* entre estos dos grupos de fenómenos es innegable, pero el caso de la recombinación es particular.

La recombinación generalizada, entendida como toda forma de intercambio de material genético entre dos individuos, dando como resultado un nuevo individuo (sea por transferencia horizontal entre bacterias o en la meiosis en organismos diploides) es sin lugar a dudas la mayor fuente de variabilidad genética. Dado que el material genético se transcribe linealmente, la ubicación dentro del genoma de un fragmento de secuencia es absolutamente fundamental para que tenga una función. Por otro lado, pero en la misma línea, la inserción aleatoria de una secuencia en medio de otra secuencia tendría intuitivamente mucha más probabilidad de tener efectos negativos que positivos. De modo que tanto para no ser destructiva como para ser constructiva, intuitivamente por lo menos, tiene sentido pensar que la recombinación está asociada a la estructura del material genético en sí.

Evidencia experimental acumulada a partir del descubrimiento de la estructura del DNA nos ha mostrado que el material genético tiene distintos niveles de organización. Esta organización, además de ser "jerárquica", por decirlo de alguna manera, también es modular, en el sentido de que por ejemplo, un codón específico codifica para un mismo aminoácido independientemente de su posición o que de una secuencia corta de aminoácidos particulares siempre se obtiene la misma forma funcional y estructura tridimensional. La modularidad, pues, está presente en todos los niveles de organización del genoma y es una característica que le ha conferido a los genomas de los seres vivos robustez, pero

también adaptabilidad.

La adaptabilidad es finalmente piedra angular de la evolución y en este trabajo se ha planteado que tanto la robustez como la adaptabilidad del genoma se consiguen mediante una recombinación estructurada en la cual se privilegia el carácter explotativo de la recombinación en etapas tempranas de la evolución (de una cadena sin función), cuando aún no se han conformado módulos funcionales y en la cual, a medida que se van conformando estos módulos, los mecanismos moleculares de la recombinación evitan recombinar en puntos intramódulo (dentro de un módulo) y lo hacen mucho más en puntos intermódulo (entre dos módulos), manifestándose así el carácter explotativo de la recombinación.

Regresando pues a lo anteriormente dicho, la recombinación es un fenómeno evolutivo particular pues se manifiesta tanto en la generación de nuevas secuencias como en la preservación de las secuencias funcionales ya conformadas. El análisis fundamental que se ha llevado en esta trabajo ha sido comparar el efecto de una recombinación aleatoria, que no discrimina entre puntos de cruza dentro o fuera de regiones funcionales o módulos y que actúa irrespectivamente de la adecuación de los mismos, y una recombinación estructurada, que hemos simulado mediante los algoritmos *Rul0* y *RulN*.

El algoritmo *RulN* comienza con una distribución homogénea en la elección de puntos de cruza y conforme se van formando módulos con alta adecuación va recombinando preferencialmente en la frontera de éstos. Nuestros resultados muestran que una población que recombina utilizando el algoritmo *RulN* tiene un mejor desempeño (es decir, que la adecuación promedio de la población aumenta más rápidamente) que una población con recombinación aleatoria permanente.

Estos resultados dependen evidentemente del paisaje de adecuación en el cual se simulen las poblaciones. Imitando el carácter modular del genoma se consideraron paisajes de adecuación independientes (aunque todos iguales) para cada bloque, es decir, que cada bloque se evolucionó en un paisaje de adecuación independiente y por otro lado, los individuos, formados por un número de bloques determinado se evolucionaron en un paisaje altamente modular (con nula epistasis).

El paisaje de adecuación DAP (doble aguja en un pajar) se eligió como el paisaje de adecuación para los bloques en la mayoría de la simulaciones y de ellas se obtuvieron la mayoría de las conclusiones interesantes de las simulaciones. Por un lado se mostró que clasificar a los paisajes DAP mediante la distancia *PPCD* (por puntos de cruza que destrozan) es un mucho mejor criterio que clasificarlos por la distancia *Hamming* entre agujas. Se mostró también que un efecto de utilizar el algoritmo *RulN* permite la *coexistencia de agujas*, es decir, la presencia en la población de dos alelos (representados cada uno por una aguja del paisaje DAP) en cada bloque, independientemente de la distancia *PPCD* entre las agujas. Se encontraron además, efectos geométricos no previstos por nuestras ecuaciones, que consisten en una correlación entre la eficiencia de la búsqueda de óptimos y la ubicación de los bloques en la cadena representativa del individuo: entre más cercanos estén los bloques a las fronteras de la cadena,

más rápidamente encuentran los óptimos.

Desde el punto de vista teórico, hemos analizado el panorama teórico en torno a la recombinación desde los algoritmos genéticos y desde la genética de poblaciones, para observar que muchas preguntas que históricamente han sido guía en biología evolutiva como si la recombinación es buena o mala, se pueden atacar con otras preguntas más relevantes dentro del contexto teórico del cómputo evolutivo, como por ejemplo, cuáles son las máscaras adecuadas de recombinación dada la conformación actual de la población. En ese sentido, el desequilibrio de ligamiento ha probado ser un medida clara e importante a la hora de determinar si la recombinación es benéfica o no para la población bajo determinadas condiciones.

A pesar de no haber abordado muchas de las preguntas planteadas de forma exhaustiva, en este trabajo se hizo un repaso histórico del estudio de la recombinación desde dos disciplinas que a pesar de haber compartido tema de estudio desde el nacimiento de ambas, en la práctica se han encontrado curiosamente alejadas. Utilizar ideas de una rama de la ciencia en otra ha sido problemente una de las grandes fuentes de descubrimientos y aplicaciones novedosas en la historia de la ciencia. Los resultados no son siempre geniales, sin embargo, esta práctica permite ampliar los horizontes de ambas disciplinas y generar un nicho donde nuevas preguntas y respuestas habrán de surgir.

Bibliografía

[1] Charlesworth B., Betancourt A. J., Kaiser V. B., and Gordo I. Genetic recombination and molecular evolution. In *Cold Spring Harb Symp Quant Biol*, 2009.

[2] N. H. Barton. Why sex and recombination? *Cold Spring Harbor Symp Quant Biol*, 74, 2009.

[3] G. Coop and M. Przeworski. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8:23–34, 2007.

[4] J. F. Crow and M. Kimura. Evolution in sexual and asexual populations. *The American Naturalist*, XCIX(909), 1965.

[5] C. Darwin. *On the Origin of Species by Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

[6] K. A. De Jong and W. M. Spears. A formal analysis of the role of multipoint crossover in genetic algorithms. *Annals of Mathematics and Artificial Intelligence*, 5:1–26, 1992. 10.1007/BF01530777.

[7] A. E. Eiben and C. A. Schippers. On evolutionary exploration and exploitation. *Fundamenta Informaticae*, 35:1–16, 1998.

[8] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 7(8):737–756, 1974.

[9] R. A. Fisher. *The genetical theory of natural selection*. Clarendon Press, 1930.

[10] D. J. Futuyma. *Evolution*. Sinauer Associates, Inc., Sunderland, MA, USA, 2005.

[11] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[12] H. Handa, N. Baba, O Katai, and T. Sawaragi. Coevolutionary genetic algorithm with effective exploration and exploitation of useful schemata. In *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems*, pages 424–427, 1997.

[13] M. Hartfield, S. P. Otto, and P. D. Keightley. The Role of Advantageous Mutations in Enhancing the Evolution of a Recombination Modifier. *Genetics*, 184(4):1153–1164, 2010.

[14] J. Hey. What's so hot about recombination hotspots? *PLoS Biology*, 2:730–733, 2004.

[15] J. H. Holland. *Adaptation in Natural and Artificial Systems*. MIT Press, 1992.

[16] Chen J., D. N. Coooper, N. Chuzhanova, C. Férec, and G. P. Patrinos. *Nature Reviews Genetics*, 8:762–775, 2007.

[17] A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, 2002.

[18] A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, Gudjonsson S. A., M. L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467:1099–1103, 2010.

[19] A. Lozano, V. Mireles, D. Monsiváis, C. R. Stephens, S. A. Alcalá, and F. Cervantes. Building blocks and search. In *MICAI*, volume 5845 of *Lecture Notes in Computer Science*, pages 704–715. Springer, 2009.

[20] J. G. March. Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1):71–87, 1991.

[21] J. Maynard Smith. *The evolution of sex*. Cambridge University Press, 1978.

[22] G. McVean. What drives recombination hotspots to repeat dna in humans? *Philosophical Transactions of the Royal Society B*, 365:1213–1218, 2010.

[23] G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584, 2004.

[24] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[25] H. J. Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932.

[26] H. J. Muller. The relation of recombination to mutational advance. *Mutation Res.*, 1:2–9, 1964.

[27] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321 – 324, 2005.

[28] K. Paigen and P. Petkov. Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, 11:221–233, 2010.

[29] R. Poli. Exact schema theory for genetic programming and variable-length genetic algorithms with one-point crossover. *Genetic Programming and Evolvable Machines*, 2(2):123–163, 2001.

[30] D. A. Rosenblueth and C. R. Stephens. An analysis of recombination in some simple landscapes. In A. H. Aguirre, R. M. Borja, and C. A. R. García, editors, *Proceedings of the 8th Mexican International Conference on Artificial Intelligence*, pages 716–727, 2009.

[31] D. A. Rosenblueth and C. R. Stephens. A theoretical analysis of the relation beween recombination and modularity. En preparación, 2011.

[32] M. P. Scott, P. Matsudaira, H. Lodish, J. Darnell, L. Zipursky, C. A. Kaiser, A. Berk, and M. Krieger. *Molecular Cell Biology*. W. H. Freeman, San Francisco, fifth edition, 2004.

[33] H. A. Simon. The architecture of complexity. In *Proceedings of the American Philosophical Society*, volume 106, pages 467–482, 1962.

[34] W. Spears. Crossover or mutation? In Whitley D., editor, *Foundations of Genetic Algorithms*, volume 2, pages 221–237. Morgan Kaufmann, 1992.

[35] C. R. Stephens and J. Cervantes. Just what are building blocks? In C. R. Stephens, M. Toussaint, D. Whitley, and P. Stadler, editors, *Foundations of Genetic Algorithms*, volume 4436 of *Lecture Notes in Computer Science*, pages 15–34. Springer Berlin / Heidelberg, 2007.

[36] C. R. Stephens and R. Poli. *Taming the Complexity of Evolutionary Dynamics: From microscopic models to schema theory and beyond*. Springer, 2011.

[37] C. R. Stephens and H. Waelbroeck. Schemata evolution and building blocks. *Evol. Comput.*, 7(2):109–124, 1999.

[38] J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. The double-strand break repair model for recombination. *Cell*, pages 25–35, 1983.

[39] A. A. R. Sá, A. O. Andrade, A. B. Soares, and S. J. Nasuto. Exploration vs. exploitation in differential evolution. In *Proceedings of the 2008 Convention on Artificial Intelligence and Simulation of Behaviour*, AISB 2008, pages 1–7, 2008.

[40] D. Thierens and D. Goldberg. Convergence models of genetic algorithm selection schemes. In Y. Davidor, H.-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature - PPSN III*, pages 119–129. Springer-Verlag, Berlin, 1994.