Biblioteca de C++ para Simulación Genética

Fabián Menéndez, Hugo Lamelas, Carlos Carleos, Jesús Á. Baro

Resumen

Presentamos una biblioteca en C++ para la simulación del proceso de trasmisión de la información genética. Su principal aplicación es el estudio del ligamiento genético, aunque se ha pretendido que su diseño resulte lo más general posible. Gracias a la programación orientada a objetos, la utilización de la biblioteca por parte de cualquier programador es sencilla y totalmente independiente de la estructura interna de la misma. Se está desarrollado asimismo un interfaz de usuario para entornos gráficos que permite administrar la realización de simulaciones concurrentes.

Introducción

Se ha elaborado una biblioteca con clases que simulan el proceso de la herencia, concretamente, el proceso de formación de gametos de un individuo (meiosis) y la generación de un nuevo individuo (apareamiento).

La idea surge de la necesidad de simular los procesos de trasmisión de la información genética en el marco del análisis de ligamiento. La sucesiva complejidad de los modelos empleados llevó a considerar la conveniencia de crear una biblioteca de "objetos genéticos" en lugar de seguir empleando algoritmos *ad hoc* en Perl. Se representaron las principales entidades biológicas involucradas en el proceso (desde alelo, locus... hasta población y especie) a través de una jerarquía de clases codificadas en C++.

La biblioteca se centra en el análisis de ligamiento, aunque no es difícil extender convenientemente las clases implementadas o añadir nuevas clases. Se han reproducido los diseños experimentales más habituales en mejora genética animal (F_2, retrocruce, familias de medio-hermanos).

Aunque la biblioteca está orientada al estudio de la segregación de genes individuales (discretos), da soporte al modelo de herencia mixta al permitir la asignación de un fondo poligénico infinitesimal, y al análisis multicarácter, permitiendo que un mismo gen pueda controlar diversos caracteres (pleiotropía).

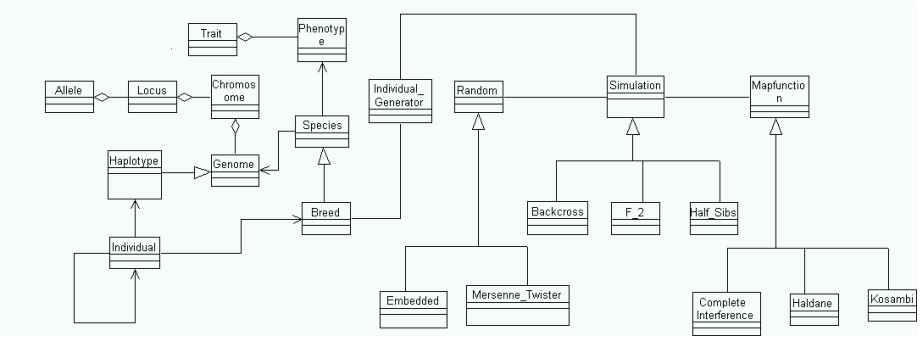
Diseño

Las entidades biológiicas representadas son: alelo, locus, cromosoma, genoma, haplotipo, carácter, fenotipo, individuo, población o raza ("breed"), especie. La clase "species" alberga el mapa genético (posición de los genes en los cromosomas

y efecto sobre los caracteres) y la clase "breed" establece las frecuencias alélicas y la varianza debida al ambiente.

El fenotipo de un individuo para cierto carácter se modela como la suma de un valor de fondo poligénico, más las contribuciones de los diferentes genes individuales que afectan al carácter, más una perturbación gaussiana debida al ambiente. Se supone que no existe interacción entre genes (epistasis), aunque sí entre alelos (dominancia). Por cada carácter que gobierna un locus pleiotrópico, se le asocia a éste una matriz simétrica $m \times m$ si el locus tiene m alelos, en la que el elemento (i,j) es el efecto del gen cuando están presentes los alelos i y j. Dado que el interés primero de la aplicación era el estudio de caracteres cuantitativos, el valor asociado a cada par de alelos es un número real, aunque el código puede adaptarse fácilmente para utilizar números difusos[1].

A partir de las clases anteriormente definidas puede llevarse a cabo la simulación de cualquier estructura genealógica. Aun así, se crearon tres clases C++ especiales para automatizar la tarea de simular los diseños experimentales más habituales: familias de medio-hermanos, retrocruces y F_2. A la hora de definir una simulación el usuario puede elegir entre una serie de generadores de números aleatorios (el propio de la arquitectura, el "Mersenne Twister"[2]), así como de funciones de mapa (Haldane, Kosambi, interferencia completa). El resultado de una clase de simulación es una lista de individuos. Se definieron objetos que resumen la información de la simulación, en forma de recuentos alélicos; el usuario puede agrupar a los individuos según su fenotipo, y solicitar recuentos alélicos en las categorías formadas. Los resultados pueden exportarse en formato propio del programa (tipo HTML), o bien en los formatos usados por Cri-Map, Genepop y Pedigree Viewer.



En la figura (diagrama OMT simplificado) se muestra la relación entre las principales clases que conforman la biblioteca.

Se pueden introducir datos para la generación de simulaciones en un archivo con formato similar al HTML. Al principio del archivo se definen los parámetros generales de la simulación (tipo y tamaño de familia, generador de aleatorios, función de mapa). A continuación se define el mapa genómico. Las frecuencias alélicas y las varianzas ambientales se fijan en sucesivas secciones, que se corresponden con las razas que intervienen en la simulación.

Discusión

Se ha empleado C++ por ser un lenguaje ampliamente extendido, para el que existen compiladores en casi todas las plataformas y que dispone de un modelo ANSI. Tiene un soporte de programación orientada a objetos (POO) muy potente y la eficiencia del código generado es muy alta. Es enlazable con otros lenguajes de programación: C, Java, Ada, Perl.

En la implementación de la biblioteca se han utilizado la mayoría de las herramientas que la POO y el C++ ofrecen para la programación, a saber: la herencia, composición, polimorfismo y ligadura dinámica, lo que dota al conjunto de gran flexibilidad y potencia .

Se ha escogido el enfoque POO con vistas a la reusabilidad, es decir, la inclusión de objetos aislados de la biblioteca en otros trabajos. Los diseños resultantes con POO son claros y elegantes, por lo que el código fuente es fácil de mantener y modificar.

Se está desarrollando un interfaz de entorno gráfico de ventanas, con estructura multihilo. Se espera ejecutar en varias plataformas (M$ Windows$^{(R)}$, Unix). Asimismo, se espera ampliar la biblioteca con el tratamiento del modelo umbral, nuevos generadores de números aletorios y la automatización de simulaciones para otros diseños experimentales.

Referencias

[1] Lamelas H, Riesgo E, 1998, "Aritmética y representación gráfica de números difusos asistidas por ordenador", Proyecto Fin de Carrera, EUITIG, Universidad de Oviedo, Gijón, Asturias.

[2] Matsumoto M, T Nishimura, 1998, Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Trans Model Comp Simul 8:3-30.