Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 2, Mes Febrero, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 18-28
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: Soluciones Informáticas
Recibido: 20/11/16 | Aceptado: 20/12/16 | Publicado: 27/02/17

Componente para análisis estadísticos de coeficientes de endogamia para la solución informática SEEGEN-R

Component for statistical analysis of inbreeding coefficients for the SEEGEN-R computing solution

Ernesto Alfredo Molina Suárez [1*], Dayana Joseph Smarth [2], Dayana Mendoza Peña [3], Bernardo Hernández González [4]

[1] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Boyeros, Ciudad de La Habana, Cuba. eamolina@uci.cu
[2] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Boyeros, Ciudad de La Habana, Cuba. djoseph@ uci.cu
[3] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Boyeros, Ciudad de La Habana, Cuba. dmendoza@ uci.cu
[4] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Boyeros, Ciudad de La Habana, Cuba. bhernandez@uci.cu

[*] Autor para correspondencia: eamolina@uci.cu

Resumen

En el Centro Nacional de Genética Médica (CNGM), se realizan los cálculos estadísticos de los estudios de Epidemiología Genética utilizando el Sistema Estadístico de Epidemiología Genética – basado en R (SEEGEN-R), herramienta desarrollada en la Universidad de las Ciencias Informáticas (UCI) que contiene módulos para la realización de estudios de Genética Poblacional, Epidemiología Tradicional y Epidemiología Genética. Actualmente, en el CNGM, se realizan análisis estadísticos de los coeficientes de endogamia de la población, pero SEEGEN-R no contiene las funciones necesarias para la realización de los mismos. El objetivo de la presente investigación es desarrollar un componente que permita la integración de los estudios de coeficientes de endogamia para su uso por los especialistas del CNGM. Se utiliza Java como lenguaje de programación para manejar la presentación de los análisis a los usuarios y R como motor de cálculo de las funciones estadísticas. En este componente se utilizó RUP como metodología de desarrollo, Netbeans como entorno de desarrollo integrado y la librería de clases RServe para

comunicar los lenguajes de programación utilizados. Como resultado se obtuvo un componente, que satisface las necesidades del CNGM, para realizar los estudios de coeficientes de endogamia en la solución informática SEEGEN-R.

Palabras clave: análisis estadísticos, coeficientes de endogamia, estadísticos F, genética médica, genética poblacional.

Abstract

The National Centre for Medical Genetic (CNGM), perform statistical calculations of the studies of genetic epidemiology using the Statistical System of Genetic Epidemiology – R based (SEEGEN-R), tool developed at the University of Informatics Sciences (UCI) that contains modules for studies of population genetics, traditional epidemiology and genetic epidemiology. Currently, in the CNGM, statistical analysis of the inbreeding coefficients in the population are made, but SEEGEN-R does not contain the functions required for the realization of the same. The objective of this research is to develop a component that allows the integration of studies of inbreeding coefficients for use by the CNGM specialists. Is used Java as a programming language to manage the presentation of the analyses to the users and R as an engine for calculation of statistical functions. RUP as methodology development, Netbeans as IDE and RServe library was used in this component to communicate the used programming languages. Was obtained as result a component, which meets the needs of the CNGM to make studies of inbreeding coefficients in the software SEEGEN-R.

Keywords*: statistical analysis, inbreeding coefficients, F-statistics, medical genetics, population genetics*

Introducción

Epidemiología Genética, disciplina de la Medicina surgida a mediados de los años 80, trata de comprender la interacción entre los factores genéticos ambientales que dan origen a las enfermedades del ser humano. Está considerada como una ciencia básica de la medicina preventiva, además de una fuente de información para la formulación de políticas de salud pública. En los últimos años ha tenido un desarrollo notable gracias a la biología molecular y el despliegue de sus marcadores genéticos, además de los avances de la tecnología en esta rama de la medicina. (Khoury MJ, 1993)

Actualmente, las iniciativas en este campo de investigación se encuentran en plena expansión, ofreciéndose diversos programas educativos y de investigación por todo el mundo. La Sociedad Internacional de Epidemiología Genética suma cada año nuevos miembros, evidenciando así la evolución de la epidemiología como ciencia.

En Cuba, la principal fortaleza de la genética radica en la introducción de los servicios de genética clínica en la atención primaria de salud, gracias a la organización vigente en la Red Nacional de Genética Médica que va desde el nivel primario (consultorios y policlínicos), hasta el terciario que son las instituciones, como el Centro Nacional de Genética Médica (CNGM), que constituye su institución rectora. Esta red tiene dentro de sus funciones principales las investigaciones básicas y aplicadas en el campo de la Genética Médica, la Inmunología, la Bioquímica y la Epidemiología Genética.

Con los avances tecnológicos y el conocimiento biológico que subyace en la acción de los genes, se puede decir que la Epidemiología Genética es una disciplina que combina el método epidemiológico con el genético, para estudiar la variación genética en poblaciones humanas y su relación con los cambios fenotípicos normales y patológicos. Permitiendo determinar la manera en que los factores de riesgos presentes en el medio ambiente interactúan con la constitución genética de una población determinada.

Un estudio importante dentro de la Epidemiología Genética es el de Genética Poblacional, que se ocupa de predecir las consecuencias que entrañan la estructura de la población y los fenómenos de selección y mutación para los fenotipos constitucionales y las enfermedades.

Dentro de la Genética Poblacional se realizan los estudios de análisis estadísticos F, que examinan los coeficientes de endogamia en una población. Al analizar estos coeficientes se pueden obtener descripciones de los niveles y el grado esperado de reducción de heterocigosidad en una población e igualmente se puede medir la correlación entre los genes extraídos en diferentes niveles de una población dividida jerárquicamente (Eguiarte, 2010).

En el CNGM, los estudios sobre análisis estadísticos de Genética Poblacional se realizan utilizando diferentes software estadísticos, los cuales no cubren muchas de las funcionalidades demandadas por los especialistas y en varios casos son herramientas propietarias. Lo que hace necesario el pago de licencias en cada una de las instituciones donde se vayan a utilizar, constituyendo un gasto considerable de dinero para el país.

Estas herramientas no son pertinentes para la realización de los estudios ya que no arrojan un resultado completo sobre los mismos, teniendo los especialistas que recurrir a cálculos manuales de fórmulas matemáticas complejas para completar la investigación iniciada. Por lo que, para obtener los resultados, se emplea una cantidad de tiempo considerable y, al depender del factor humano para realizar complejas ecuaciones, se pueden cometer errores de cálculo que surgen al trabajar con juegos de datos grandes como los que se almacenan en el CNGM.

Actualmente se invierte mucho tiempo de la investigación al comprobar los cálculos realizados varias veces para no arribar a conclusiones equívocas sobre los estudios de Genética Poblacional del país (Marcheco, 2009). Una investigación con resultados inexactos puede traer consigo una mala inversión de dinero o el desarrollo de tecnologías innecesarias para resolver padecimientos de la población.

Partiendo de la problemática planteada se propone como objetivo de la investigación desarrollar una solución informática capaz de realizar análisis estadísticos en estudios de Epidemiología Genética.

Materiales y métodos

Se propone desarrollar un sistema capaz de realizar análisis estadísticos en estudios de Epidemiología Genética. Esta solución estará basada en componentes independientes, destinados a los estudios de Genética Poblacional, la Epidemiología Genética y la Epidemiología Tradicional. Con el desarrollo constante de la Genética Médica y la aparición de nuevos estudios y métodos se hace necesario ir adicionando componentes a la aplicación que permitan aumentar los análisis a realizar, para que la solución informática agrupe y centralice la mayoría de los análisis sostenidos en la esfera científica y lograr eliminar los problemas que presenta el CNGM, contribuyendo de esta manera a la salud de la sociedad actual y futura.

Diseño del sistema

Se le llama actor a toda entidad externa al sistema que guarda una relación con éste y que le demanda una funcionalidad. Esto incluye a los operadores humanos pero también incluye a todos los sistemas externos, además de entidades abstractas, como el tiempo. Es frecuente encontrar sistemas que deben efectuar operaciones automáticas en determinados momentos, en estos casos se asume que esa operación automática está condicionada también por el

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 2, Mes Febrero, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 18-28
http://publicaciones.uci.cu

tiempo de ahí que se concluya que en estos casos el actor es el tiempo o el propio sistema (Rumbaugh et al., 1999). En la Tabla 1 se muestran el actor del sistema y sus funciones correspondientes:

Tabla 1. Actores del sistema

Actor	Descripción
Especialista en Genética	Interactúa con el sistema y es el encargado de realizar los análisis y estudios de estadísticos F.

Un Caso de Uso (CU) del sistema es una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento iniciado por un actor principal. Los diagramas de casos de uso sirven para especificar la comunicación y el comportamiento de un sistema mediante su interacción con los usuarios y/u otros sistemas (Landacay, 2008).

Con el objetivo de ofrecer un mayor entendimiento del problema en la figura 1 se muestra el diagrama de CU correspondiente a la propuesta de solución, donde cada CU responde a una interfaz en el componente.

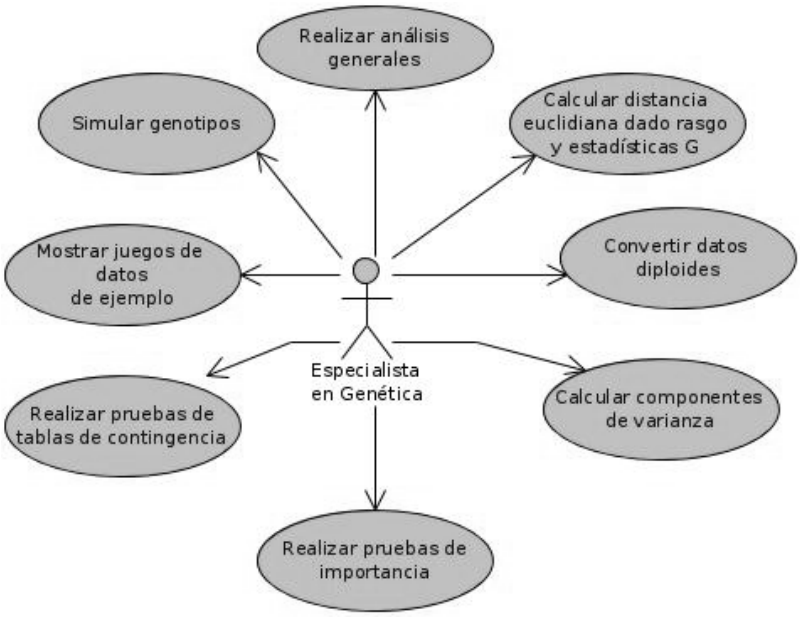

Figura 1. Diagrama de Casos de Uso del Sistema

Herramientas y tecnologías

Una vez analizadas las tendencias actuales en el desarrollo de software se seleccionó RUP como metodología para guiar el proceso de desarrollo. Del estudio de las herramientas y tecnologías se definió la línea base de la arquitectura utilizando: UML como lenguaje de modelado, Visual Paradigm 8.0 como herramienta CASE, Java 1.6 y R 3.2.0 como lenguajes de programación; siendo NetBeans 8.0 el entorno de desarrollo integrado a utilizar.

Arquitectura del sistema

Una arquitectura de software se utiliza para estructurar y guiar el desarrollo de un sistema, con lo cual se puede satisfacer los atributos de calidad (Figueroa, 2007). Este proceso se refiere a las estructuras compuestas de elementos con propiedades visibles de forma externa y las relaciones que existen entre ellos. Juega un papel fundamental en las etapas del desarrollo de un software permitiendo tomar las decisiones críticas del sistema al principio de su concepción. (Addison, 2003)

La aplicación SEEGEN-R usa una arquitectura basada en componentes, debido a esto le permite el desarrollo de componentes de forma independiente y luego integrarlos (Kuchana, 2004). Dentro de sus componentes presenta:

Módulo Base: es el encargado de cargar los datos genéticos para realizar cada uno de los estudios ya sea desde un fichero guardado o creados en la misma aplicación por el genetista, también cumple con la función de mostrar en el menú correspondiente cada una de las vistas a mostrar para los especialistas y por último recibe los datos de los resultados y los muestra en pantalla.

SeegenrApiPlugin: es la biblioteca de clases encargada de conectar el módulo base con los componentes desarrollados, brindando una interfaz de conexión entre estos que permite el intercambio de información. Comunica al módulo base el menú que debe crear para acceder a las vistas del componente y los resultados de los estudios realizados. Comunica a los componentes integrados el idioma del sistema para visualizar los textos correspondientes y brinda los datos cargados previamente en el módulo base.

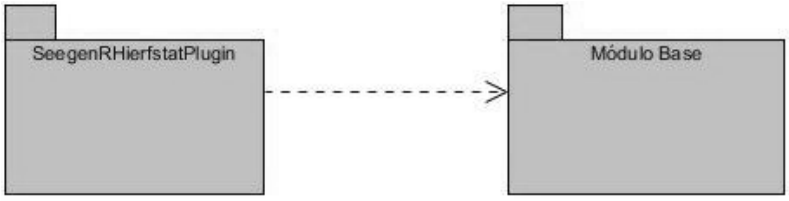

Figura 2. Diagrama de paquetes

Estándares de codificación

Un estándar de codificación comprende todos los aspectos a tener en cuenta en la generación de código para que este sea legible y asegurarse que todos los programadores del proyecto trabajen de forma coordinada (Hernández et al., 2015). Usar técnicas de codificación sólidas y realizar buenas prácticas de programación con vistas a generar un código de alta calidad es de gran importancia para la calidad del software y para obtener un buen rendimiento (Microsoft, 2003).

En el desarrollo del componente los estándares a seguir están regidos por los definidos para el desarrollo de la solución informática SEEGEN-R, algunos de los cuales se presentan a continuación:

- Los nombres de las clases deben ser sustantivos, cuando son compuestos tendrán la primera letra de cada palabra que lo forma en mayúsculas, manteniendo los nombres de las clases simples y descriptivos.
- En los comentarios de las clases debe aparecer el autor de esta y el objetivo de la misma.
- Inicializar las variables locales donde se declaran. La única razón para no inicializar una variable donde se declara es si el valor inicial depende de algunos cálculos que deben ocurrir.
- Utilizar nombres en plural para arreglos, tipos de datos abstractos TDA o matrices de objetos.
- Todos los métodos y clases deben estar comentariadas. Los comentarios deben ser añadidos de forma que resulten prácticos, para explicar el flujo del código y el propósito de las funciones o variables.
- Las líneas en blanco mejoran la facilidad de lectura separando secciones de código que están lógicamente relacionadas por lo que se debe usar siempre una línea en blanco entre métodos y bloques de código.

Resultados y discusión

Pruebas del sistema

Las aplicaciones, en general cualquier mecanismo diseñado e implementado por un humano, son propensas a tener fallos, surge por tanto la necesidad de asegurar en lo posible, la calidad del producto. El único instrumento adecuado para determinar el status de la calidad del mismo es el proceso de pruebas. En este proceso se ejecutan pruebas dirigidas a componentes del software o al sistema de software en su totalidad, con el objetivo de medir el grado en que el software cumple con los requerimientos y presentar información sobre la calidad del producto a las personas responsables de éste (). Los diferentes niveles de prueba se presentan a continuación:

- Prueba de Sistema: Se hace para verificar el programa final, cuando el sistema funciona como un todo y cada uno de los componentes de hardware y software están integrados en su totalidad.
- Prueba de Integración: Se realiza para asegurar que los componentes que son combinados para ejecutar un CU funcionen correctamente (Pressman, 2010).

Pruebas de integración descendente

Para comprobar el correcto funcionamiento del componente es necesario realizar las pruebas de integración con el módulo base de la solución informática SEEGEN-R, para esto se realizan las pruebas de integración descendente en la cual se integran los módulos moviéndose hacia abajo por la jerarquía de control del sistema, comenzando por el módulo de control principal (módulo base). Los módulos subordinados al módulo de control principal se van incorporando en la estructura.

Para ejecutar las pruebas de integración descendente se parte del módulo base de la solución informática SEEGEN-R y se van agregando los componentes a integrar, luego se ejecuta la aplicación principal, la cual identifica el nuevo componente y lo integra al menú principal del sistema. Se accede a las opciones del componente que se desean probar, de ocurrir algún fallo en la integración de una funcionalidad, se accede al registro de errores de la aplicación y se pueden observar los mensajes ocurridos durante la ejecución del sistema.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 2, Mes Febrero, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 18-28
http://publicaciones.uci.cu

Resultados de la aplicación de las pruebas

Como parte de la ejecución de las pruebas de caja negra se realizaron 3 iteraciones de pruebas, representadas en la figura 3. En la primera iteración se detectaron 15 no conformidades, clasificadas en 12 no significativas y 3 significativas. Una vez corregidas, se procedió a realizar la segunda iteración donde se detectaron 5 no conformidades, de las cuales 4 fueron significativas y 1 no significativa. Por último, se realizó una tercera iteración en la cual se detectaron 0 no conformidades.

Para la generación de los casos de prueba de sistema y de pruebas de integración descendente se utilizará la técnica de partición equivalente del método de caja negra. Con el fin de llevar a cabo la evaluación de la propuesta de solución actual, atendiendo a las características y objetivo del sistema, se seleccionó la aplicación de dos tipos de prueba correspondientes a las pruebas a nivel de componentes.

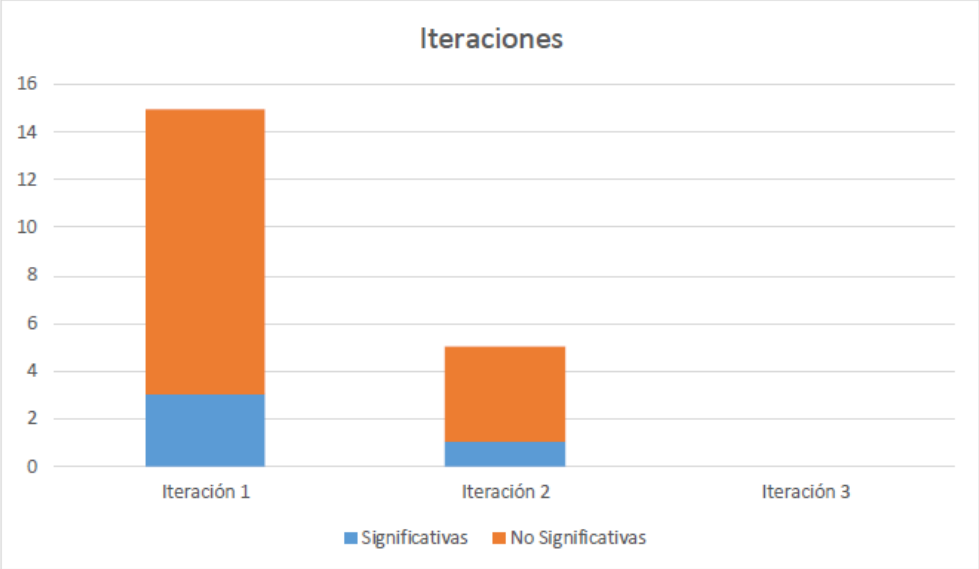

Figura 3. Resultados de las iteraciones de las pruebas

Conclusiones

Con el análisis de los estudios de coeficientes de endogamia que se realizan en el CNGM se logró una correcta identificación de los requisitos funcionales y no funcionales del componente.

El estudio de las aplicaciones que permiten realizar los cálculos estadísticos de genética poblacional permitió la selección de la solución informática SEEGEN-R como herramienta para desarrollar la solución propuesta.

A partir de la realización del análisis de la solución informática SEEGEN-R se seleccionaron las tecnologías, herramientas y metodología de desarrollo que se ajustan a la situación del negocio. Se realizó un correcto diseño de las clases del sistema donde se obtuvieron los diagramas de clases del diseño y de secuencia, haciendo un buen uso de los patrones de diseño sobre la base del patrón arquitectónico Dos-capas.

El componente brinda un conjunto de nuevas funcionalidades a la solución informática SEEGEN-R, permitiendo el cálculo de coeficientes de endogamia y de esta forma se logra una mayor integración de los estudios que se realizan en el CNGM; dando cumplimiento de lo establecido en la especificación de requisitos de software.

La realización de pruebas de sistema y de integración así como la resolución de las no conformidades encontradas demostraron el correcto funcionamiento del componente implementado.

Referencias

ADDISON W., L. BASS, P. CLEMENTS, R. KAZMAN. Software Architecture in Practice, 2nd Edition. 2003.

EGUIARTE, L. Flujo génico, diferenciación y estructura genética de las poblaciones, con ejemplos en especies de plantas mexicanas. México. s.n., 2010.

FIGUEROA, A. Descripción de la Arquitectura del Sistema. Entorno de Simulación Robótico. Universidad de la Republica Montevideo. Uruguay: s.n., 2007.

HERNÁNDEZ, B. RODRÍGUEZ, O. MAR, O. Sistema para la gestión de impresiones del Vicedecanato de Administración de la Facultad 6 de la Universidad de las Ciencias Informáticas. 2015.

KHOURY MJ, BEATY TH, COHEN BH. Fundamentals of genetic epidemiology. New York: Oxford University Press: s.n., 1993.

KUCHANA, PARTHA. Software Architecture Desing Patterns in Java. 2004.

LANDACAY, KATTY. UML: Caso de uso. [En línea] 2008. [Citado el: 20 de Enero de 2015.] http://es.slideshare.net/ktyk/uml-casos-de-uso.

MARCHECO, TERUEL BEATRIZ. El Programa Nacional de Diagnóstico, Manejo y Prevención de Enfermedades Genéticas y Defectos Congénitos de Cuba: 1981-2009. [En línea] 2009. [Citado el: 4 de Abril de 2015.] http://bvs.sld.cu/revistas/rcgc/v3n2_3/rcgc1623010%20esp.htm.

MESA, DRA. MARIA SOLEDAD. Departamento de Zoología y Antropología Física – UCM. [En línea] 2010. [Citado el: 4 de Abril de 2015.] http://pendientedemigracion.ucm.es/info/antropo/genetica.htm.

MICROSOFT. Revisiones de código y estándares de codificación. [En línea] https://msdn.microsoft.com/es-es/library/aa291591%28v=vs.71%29.aspx. 2003

MOZO, MARIA VICTORIA. Biology-Online Dictionary. [En línea] 2011. https://www.biology-online.org/dictionary/Genetic_locus.

PARDO C. AND GARCIA F. Diagrama de Clases en UML 1.1. 1998.

PRESSMAN, R. S. Ingeniería del software Un enfoque práctico. Edtion ed., 2010. 805 p. ISBN 978-607-15-0314-5.

RUMBAUGH, J, JACOBSON, I Y BOOCH, G.UML Reference Manual. 1999.

SAAVEDRA, JORGE. Lenguajes de programación. [En línea] 2008. https://jorgesaavedra.wordpress.com/2007/05/05/lenguajes-de-programacion/.