SEEGEN-R: Sistema Estadístico de Epidemiología Genética

Roberto Lardoeyt Ferrer
Centro Nacional de Genética Médica, La Habana, Cuba, lardgen@infomed.sld.cu

Yudiel La Rosa González
Universidad de las Ciencias Informáticas, La Habana, Cuba, ylarosag@uci.cu

Dayana Joseph Smarth
Universidad de las Ciencias Informáticas, La Habana, Cuba, djoseph@uci.cu

Yunier Emilio Tejeda Rodríguez
Universidad de las Ciencias Informáticas, La Habana, Cuba, yuniere@uci.cu

Darwis Rodríguez Licea
Universidad de las Ciencias Informáticas, La Habana, Cuba, drlicea@estudiantes.uci.cu

Asiel Leal Celdeiro
Universidad de las Ciencias Informáticas, La Habana, Cuba, aceldeiro@estudiantes.uci.cu

RESUMEN

La epidemiología genética es una disciplina que estudia la interacción entre los factores genéticos y ambientales que dan origen a las enfermedades del ser humano, valiéndose de marcadores genéticos, complejos algoritmos y amplias bases de datos. En estos estudios se utilizan diferentes análisis estadísticos que requieren mucho tiempo, siendo necesario el empleo de herramientas de cálculo que garanticen el procesamiento rápido y efectivo de los datos. En la actualidad los especialistas necesitan combinar diferentes soluciones informáticas para realizar los análisis necesarios para estos estudios haciendo aún más complicada su realización, comprensión y en muchas ocasiones provocando una pérdida considerable del flujo de trabajo comprometiendo la veracidad de los resultados obtenidos. La construcción de un software que agrupe los estudios de este campo le brindaría una mayor fluidez en la realización de estos estudios y siendo este el objetivo del Sistema Estadístico de Epidemiología Genética basado en R. Esta solución permite la realización de diferentes estudios de la Epidemiología Genética, Epidemiología Tradicional y la Genética Poblacional, además de facilitar tareas útiles como la carga y exportación de datos desde diferentes fuentes, graficar y salvar el trabajo realizado utilizando métodos de cifrado.

Palabras claves: epidemiología genética; genética poblacional; análisis estadísticos; lenguaje R

ABSTRACT

Genetic Epidemiology is a discipline which studies interaction among genetic and environmental factors that give rise to human diseases, using genetic markers, complex algorithms and wide databases. In these studies are used different complex statistical analysis which require of time mostly, being necessary to employ calculation tools that guarantee the fast and effective processing of data. Nowadays specialists need to combine different computing solutions to perform the necessary analysis to carry out this studies, doing their performance and understanding more complex even and in many cases, causing a considerable lost in the workflow, compromising the veracity of obtained results. The building of a software that groups studies of the Genetic Epidemiology field would provide a greater fluidity in the realization of these studies and that is precisely the main goal of Statistical System for Genetic Epidemiology based on R, whose development was guided by the methodology of the Unified Process for Software Development. This solution allows the performance of different studies of Genetic Epidemiology, Traditional Epidemiology and Population Genetics, besides facilitating useful tasks such loading and exportation of data from different

sources, to graph and to save the done work using ciphering methods.

Keywords: genetic epidemiology; population genetics; statistical analysis; R language

1. INTRODUCCIÓN

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. (Wyszynski, 1998)

Dado el drástico descenso en el coste de genotipado, los estudios de epidemiología genética suelen disponer de una gran cantidad de información y para realizar el análisis estadístico. En el Centro Nacional de Genética Médica (CNGM) de Cuba, los estudios de diseños de la epidemiología genética se realizan utilizando paquetes estadísticos como el SPSS, InfoStats, Statistics, Epidat, entre otros, los cuales no cubren todas las necesidades demandadas por los especialistas y la mayoría son herramientas propietarias, requiriendo el pago de licencias constituyendo un gasto considerable para el país.

Es significativo señalar que en nuestro país no se cuenta con una solución que brinde los más disímiles diseños descriptivos, analíticos y experimentales, y que posibilite enfocar investigaciones en esta rama científica cada vez más compleja. En este sentido se conformó un equipo multidisciplinario de investigación integrado por epidemiólogos, genétistas, matemáticos e informáticos del Centro Nacional de Genética Médica y de la Universidad de Ciencias Informáticas este equipo desarrollo el Sistema Estadístico de Epidemiología Genética - basado en R (SEEGEN-R), que agrupa los estudios de este campo brindando una mayor fluidez en la realización de sus análisis.

2. RESULTADO Y DISCUSIÓN

Se utiliza el lenguaje de programación R para el desarrollo de la capa de análisis estadístico por ser un lenguaje y un entorno en el que se aplican las técnicas estadísticas. Constituye uno de los lenguajes estadísticos más utilizados y con una amplia comunidad científica. También se utiliza Java como lenguaje de programación para la capa de presentación por la robustez y consolidación que ha tenido para la realización de software con calidad y eficiencia.

El sistema SEEGEN-R que en su primera versión cuenta de un sistema base y tres grandes extensiones: estudios de epidemiología genética, estudios de epidemiología tradicional y estudios de genética poblacional.

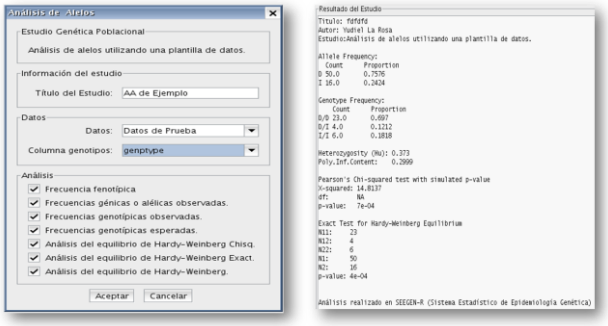

Figure 1: Interfaz de análisis de alelos y del resultado.

3. CONCLUSIONES

El sistema SEEGEN-R permite que los especialistas tengan mayor control de las enfermedades en las poblaciones y de esta forma poder tomar medidas en beneficio de las personas que las conforman.

La arquitectura basada en extensiones de SEEGEN-R permite dejar abierto el alcance del sistema, brindando la posibilidad de mejorar las extensiones existentes o de desarrollar nuevas extensiones para otros estudios.

REFERENCIAS

WYSZYNSKI, Diego F. (1998). "La Epidemiología Genética: disciplina científica en expansion". Revista Panamericana de Salud Pública, Vol. Public Health, pp 26-27.

Hey J. and WakeleyJ. (1997). "A coalescent estimator of the population recombination rate". PubMed Central, Vol. Genetics, pp 833–846.

R Foundation. The R Project for Statistical Computing, http://www.r-project.org, 24/01/2014. (fecha de consulta)

Oracle Corporation. Java programming language, http://www.oracle.com/lad/technologies/java/overview/index.html, 15/01/2014. (fecha de consulta)

Authorization and Disclaimer