Modelo conceptual bioinformático aplicado al análisis genómico de las enfermedades cardiovasculares

O. Coltell[a], D. Corella[b], J.T. Sánchez[a], R. Chalmeta[a] y J.M. Ordovás[c]

[a]Grupo de Integración y Re-Ingeniería de Sistemas. Departamento de Lenguajes y Sistemas Informáticos. Universitat Jaume I. Castellón. España.
[b]Unidad de Epidemiología y Genética Molecular. Departamento de Medicina Preventiva y Salud Pública. Universitat de València. Valencia. España.
[c]Nutrition and Genomics Laboratory. JM-USDA-Human Nutrition Research Center on Aging at Tufts University, Boston. Massachusets. Estados Unidos.

Fundamento. El análisis genómico actualmente necesita de la concurrencia de disciplinas muy alejadas del campo de la biología molecular, entre ellas la ciencia de la computación, base de la bioinformática. Las herramientas bioinformáticas trabajan fundamentalmente con modelos abstractos de la información genómica. Este trabajo tiene por objetivo la obtención de un modelo conceptual de la información genómica cuya finalidad es la construcción de sistemas de *software* bioinformáticos para el análisis de las enfermedades cardiovasculares.

Métodos. Se han aplicado 2 enfoques complementarios en ingeniería del *software*: el paradigma de la orientación a objetos, aplicando el método unificado de desarrollo de *software* y el lenguaje de modelado *unified modelling language* (UML; http://www.uml.org, http://www.omg.org), así como los sistemas multiagente, según la inteligencia artificial distribuida.

Resultados. Considerando que tanto el gen como el genotipo o el fenotipo, entre otros, son entidades objetuales a modelar, se ha construido un modelo orientado a objetos para la gestión de la información genómica y cardiovascular. En este modelo se consideran tanto los elementos de información y las relaciones entre ellos (la visión estructural), como su comportamiento en el sistema en funcionamiento (la visión dinámica). Para modelar la interacción entre los distintos factores (gen-gen y gen-ambiente) se ha considerado un modelo multiagente, donde cada factor está representado por un agente elemental. También se ha construido un pequeño prototipo correspondiente al subsistema de gestión de la información.

Conclusiones. Se ha obtenido un modelo conceptual para la construcción de herramientas bioinformáticas que apoyan al análisis genómico en las enfermedades cardiovasculares. El modelo presenta 2 vertientes: el enfoque orientado a objetos con UML para la arquitectura básica del sistema y el enfoque orientado a agentes para tratar las interacciones entre los factores genéticos y ambientales. El modelo es parte de un proyecto de mayor envergadura que pretende construir un sistema funcional completo denominado A-Genes.

Palabras clave:
Sistema bioinformático. Arquitecturas multiagente. Análisis genómico. Enfermedades cardiovasculares. Riesgo cardiovascular.

Este trabajo está financiado por el Ministerio de Educación y Ciencia de España, becas PR2002-0116 (O. Coltell) y PR2002-0115 (D. Corella) y parcialmente financiado por MCYT (DPI 2003-02515).

Correspondencia: Dr. O. Coltell Simón.
Departamento de Lenguajes y Sistemas Informáticos.
Universitat Jaume I.
Campus de Riu Sec, s/n. 12071 Castellón. España.
Correo electrónico: coltell@lsi.uji.es

Recibido el 6 de febrero de 2003 y aceptado el 6 de mayo de 2003.

BIOINFORMATIC CONCEPTUAL MODEL APPLIED TO GENOMIC ANALYSIS IN CARDIOVASCULAR DISEASES

Background. Genomic analysis requires the partnership of disciplines far removed from the field of molecular biology, one of which is

computer science, the basis of Bioinformatics. Bioinformatic tools work mainly with abstract models of genomic data. This work aimed to obtain a conceptual model of genomic data for building bioinformatic software systems for cardiovascular diseases.

Methods. Two complementary approaches in software engineering were applied: the object-oriented paradigm, using the unified method for software development and the unified modelling language (UML) (http://www.uml.org, http://www.omg.org); and multi-agent systems following distributed artificial intelligence.

Results. Considering that gene, genotype, phenotype, etc. are object entities for modelling, an object-oriented model was constructed for genomic and cardiovascular data management. This model includes information elements and their inter-relationships (the structural view) and their behaviour in the running system (the dynamic view). In order to assess interactions between different factors (gene-gene and gene-environment), a multi-agent model was considered in which each factor was represented by an elemental agent. In addition a small prototype corresponding to the data management subsystem was built.

Conclusions. A conceptual model for bioinformatic tools supporting genomic analysis in cardiovascular diseases was obtained. The model shows two aspects: the object-oriented approach with UML for the basic system architecture and the agent-oriented approach for dealing with interactions between genetic and environmental factors. The model is part of a very large project for constructing a whole functional system named A-Genes.

Key words:
Bioinformatic system. Multi-agent architectures. Genomic analysis. Cardiovascular diseases. Cardiovascular risk.

Introducción

El análisis genómico en la actualidad requiere la concurrencia de disciplinas muy alejadas del campo de la biología molecular, entre ellas la ciencia de la computación, base de la bioinformática[1,2]. Las herramientas bioinformáticas trabajan fundamentalmente con modelos abstractos de la información genómica obtenidos a partir de estudios de biología molecular, que proporcionan información de bajo nivel[3,4], y estudios epidemiológicos, que proporcionan información de alto nivel[5]. En concreto, uno de los problemas pendientes de resolver con eficiencia es el cálculo del riesgo cardiovascular o CHDR (*cardio-heart disease risk* o riesgo de padecer enfermedades cardiovasculares [ECV]) a escala individual[6-8].

En la actualidad ya existen sistemas que calculan el CHDR, pero solamente tienen en cuenta las variables bioquímicas principales y un reducido grupo de variables clínicas, tal y como se empezó a realizar en el Framingham Heart Study[9]. Estos sistemas no hacen más que aplicar un enfoque epidemiológico estadístico[10]. Es decir, la obtención de los riesgos cardiovasculares relativo y total se hace mediante la realización de estudios sobre muestras amplias de la población y en la extracción de conclusiones sobre los resultados obtenidos de los análisis estadísticos aplicados sobre los datos[5]. Estos estudios epidemiológicos son de tipo cohorte (extendidos a lo largo del tiempo y con varias generaciones de las mismas familias) o casos y controles (transversales en el tiempo y con individuos sin relación familiar). Pero la naturaleza multifactorial de las ECV obliga a tener en cuenta un conjunto más amplio de factores con valores de cada individuo.

Sin embargo, debido a que los métodos analíticos clásicos no son suficientemente resolutivos o los estudios adecuados exigen unos tamaños de muestra imposibles de alcanzar, se han buscado aplicaciones de nuevas tecnologías para abordar este problema y otros similares. Las líneas principales son la minería de datos *(data warehousing)*[11], nuevos enfoques estadísticos[12] y la aplicación del análisis de *microarrays* de ADN como tecnología de base[13,14].

Este trabajo pretende aportar otro enfoque alternativo a los mencionados mediante la aplicación de las arquitecturas multiagente y los modelos conceptuales y tecnologías basadas en el concepto de agente, tanto desde el punto de vista de la inteligencia artificial distribuida[15], como el de la ingeniería del *software* orientado a agentes[16].

El objetivo de este trabajo es la descripción de un proyecto en curso que consiste en el modelado y construcción de un sistema bioinformático, basado en el concepto de agente *software* y las arquitecturas orientadas a agentes, que se pueda utilizar como herramienta de soporte en el análisis genómico y estudios epidemiológicos en las ECV. En dicho sistema, dado un individuo y un conjunto de informaciones de ese individuo (genotipo, estilo de vida, parámetros biológicos y clínicos, etc.) que provienen de distintas fuentes y otro conjunto de informaciones que provienen de estudios de pobla-

ción, deberá ser posible calcular el CHDR de que el individuo analizado desarrolle una enfermedad cardiovascular en un plazo de tiempo determinado. Esta información será válida para decidir acerca de las medidas preventivas a aplicar a cada individuo en particular.

Los resultados presentados son una de las aportaciones de un proyecto de investigación en curso relativo a la aplicación de la bioinformática (en sus facetas de inteligencia artificial e ingeniería del *software*) en la epidemiología genómica de las enfermedades cardiovasculares. Este proyecto es fruto de una colaboración tripartita entre la Tufts University, Nutrición y Genética, la Universidad de Valencia, Epidemiología Genética y la Universidad Jaume I, Bioinformática desde 1997[17].

En la siguiente sección se describe la metodología empleada en el desarrollo del trabajo; en la tercera se muestran los resultados obtenidos, que son parciales dado que se trata de un proyecto en curso, y en la cuarta se exponen las conclusiones recogidas a lo largo de la realización del trabajo.

Métodos

Metodología y entornos de desarrollo

El enfoque utilizado para el diseño del sistema ha sido el de sistema multiagente (MAS)[15]. Se ha aplicado una metodología de desarrollo que es el método unificado de desarrollo de *software* de Rational Corp.[18]. Esta metodología utiliza como lenguaje de representación y modelado UML (*unified modeling language* de Rational Corp.)[19] y como modelo de desarrollo, el modelo iterativo incremental.

La herramienta de modelado aplicada ha sido Rational Rose Enterprise 2002 Edition, también de Rational Corp. El entorno elegido para el modelado de agentes es Agent Builder Pro 1.3a de Reticular Systems, que produce dichos agentes en módulos escritos en lenguaje Java.

El entorno elegido para la interfaz local es Borland Delphi, versión 7.0 (Object Pascal) debido a sus mejores características de versatilidad y rendimiento que los entornos Java. Y como lenguaje de integración del sistema MAS y de la interfaz se ha elegido Python. Todo ello se está desarrollando en una plataforma PC. Sin embargo, dado que todos los entornos y lenguajes mencionados tienen sus respectivas versiones para Linux (p. ej., Borland Killix es el correspondiente a Borland Delphi), no se ha descartado implementar una versión posterior para este sistema operativo.

Modelos de riesgo

El modelo para el cálculo de riesgo que se ha aplicado corresponde a la situación en que se conoce la susceptibilidad genética y se dispone de un indicador o medida actual[5]. Hay varios tipos de métodos para obtener los cálculos de riesgo. En este caso se ha considerado el método del "estudio de casos y controles con controles no emparentados", dado el perfil de la muestra donde no se incluyen individuos emparentados en los estudios habituales, aunque también se toman los datos del parentesco.

En el estudio de casos y controles con controles no emparentados se utiliza como grupo de referencia a los individuos que no manifiestan susceptibilidad genética (su genotipo es el normal sin mutaciones). Se estiman las *odds ratio* (OR) para el resto de grupos y se ajusta por las posibles variables de confusión mediante análisis de estratificación o multivariado (tabla 1). Una *odds* es la razón entre la probabilidad de la ocurrencia de un suceso y la probabilidad de la ocurrencia de su complementario. Más concretamente, una OR es la razón entre dos *odds*: la obtenida por el hecho de pertenecer a un grupo respecto a la de no pertenecer. El modelo estadístico a partir del que se obtienen estos estimadores de riesgo es la regresión logística, aunque la OR también puede calcularse a partir de las probabilidades que se obtienen aplicando el modelo loglineal multinomial[20] o el modelo loglineal de Poisson[21].

Este método permite determinar la interacción entre factores de riesgo no específicos y, posteriormente, analizar la interacción gen-ambiente. Las frecuencias de los distintos factores que se deben incluir en el estudio determinan principalmente el tamaño de la muestra[22].

En la tabla 2 se muestran los factores principales analizados agrupados por categorías: no modificables y modificables. También se incluye el grado de incidencia de cada factor en el cálculo: binario, categórico, continuo y multivariado. Un factor binario expresa su ausencia o presencia en el sistema de cálculo. Un factor categórico expresa determinados grados de incidencia discretos. Un factor continuo incide de forma continua en valores naturales, enteros, racionales o reales. Y un factor multivariado se asocia con los distintos polimorfismos de los marcadores genéticos, donde, dado un polimorfismo, se indica si existe o no la mutación (binario), y esto para todos los polimorfismos del mismo marcador y de otros que se asocian a la misma funcionalidad genética de riesgo (multivariado).

Tabla 1. Variantes de los modelos de riesgo

Modelo de riesgo multiplicativo	Modelo de riesgo aditivo
$OR_{interacción} = \frac{OR_{exposición,\ genotipo}}{OR_{exposición} \times OR_{genotipo}}$	$OR_{interacción} = \frac{OR_{exposición,\ genotipo}}{OR_{exposición} + OR_{genotipo} - 1}$
$OR_{interacción} = OR_{exposición}$: *odds ratio* del efecto de la exposición al factor ambiental aislado $OR_{genotipo}$: *odds ratio* del efecto del genotipo aislado $OR_{exposición,genotipo}$: *odds ratio* del efecto combinado del genotipo y de la exposición al factor ambiental	Interpretación de $OR_{interacción}$: $OR_{interacción}$: > 1 efecto de aumento de origen = 1 no modifica el riesgo < 1 efecto protector

[a]Para que un riesgo sea estadísticamente significativo, el intervalo de confianza del 95% no debe incluir el valor nulo de riesgo (*odds ratio* [OR] = 1).

El valor del grado de incidencia determinará la naturaleza matemática de cada uno de los factores a la hora de efectuar el cálculo del riesgo.

Resultados

El sistema se ha planteado con los requisitos siguientes:

1. Que sea modular, de forma que esté compuesto por varios subsistemas, cada uno de ellos teniendo asignada una o varias funciones principales e independientes.

2. Función principal 1: cálculo del riesgo a partir de la influencia e interacción de los distintos factores que componen el genotipo y el fenotipo de cada individuo, sin despreciar los factores de corrección poblacionales. Para ello se debe diseñar un modelo que tenga en cuenta todas las interacciones y los signos de los factores: el signo positivo indica protección y el negativo, riesgo.

3. Función principal 2: gestión del flujo de control de los módulos o subsistemas mediante un modelo de control jerárquico que combine distintos enfoques computacionales: reglas, algoritmos, etc.

4. Función principal 3: gestión y almacenamiento de toda la información que maneja el sistema: datos de los individuos, datos de los distintos estudios en que se encuadren los individuos, datos genómicos, datos bioquímicos, etc. También se debe controlar la caducidad temporal de ciertos datos y pedir su actualización.

5. Función principal 4: adquisición de las órdenes del usuario, normalmente un especialista médico, que utiliza el sistema en un acto profesional.

6. Función principal 5: presentación de resultados al usuario en forma de datos numéricos o ayudas a la decisión.

7. Función principal 6: adquisición de información en fuentes externas como Medline, OVID, bases de datos de SNP y cuantas fuentes remotas puedan ser necesarias para mantener el repositorio actualizado de forma automática o automáticamente asistida.

Según estos requisitos y otros de menor nivel, el modelo funcional preliminar obtenido se compone de 4 subsistemas que asumen las 6 grandes funciones mencionadas (fig. 1).

Por otra parte, se ha pensado que cada uno de los subsistemas puede aprovechar las características especiales que ofrecen los sistemas multiagente para el manejo de entornos heterogéneos. De esta forma, cada subsistema puede ser un sistema multiagente con tecnologías y enfoques operativos adaptados a las funciones que tiene asignadas. Por ejemplo, el subsistema de interfaz puede aprovechar la tecnología de motores de búsqueda y agentes de Internet que ya están desarrollados y disponibles.

Tabla 2. Factores de riesgo agrupados por categorías

Categoría	Factor de riesgo	Grado de incidencia
No modificables	Edad avanzada	Continuo
	Sexo masculino	Binario
	Raza	Categórico
	Antecedentes familiares	Binario
	Marcadores genéticos	Binario/ multivariado
	Apolipoproteína E	Continuo
	Lipoproteinlipasa	Continuo
Modificables	Lípidos plasmáticos	Continuo
	Tabaquismo	Categórico
	Hipertensión	Continuo
	Sedentarismo	Binario
Otros	Diabetes	Binario
	Obesidad	Binario
	Hormonas y anticonceptivos orales	Categórico
	Estrés	Binario

Proceso de desarrollo

El proyecto para la obtención del sistema completo y en funcionamiento es de tal envergadura que ha sido necesario dividirlo en subproyectos más manejables que se van realizando concurrentemente y con desfase temporal. Los subproyectos más importantes son los siguientes:

– Modelado conceptual del sistema completo.
– Modelado de diseño y prototipificación funcional de parte del sistema.
– Modelado de implementación del subsistema de riesgo.
– Modelado de implementación del subsistema de control.
– Modelado de implementación del subsistema de repositorio.
– Modelado de implementación del subsistema de interfaz.
– Integración de prototipos de los subsistemas.
– Pruebas del sistema.

Algunos de estos subproyectos son interdependientes. En esta sección se presentan los resultados de los 2 primeros subproyectos, teniendo en cuenta

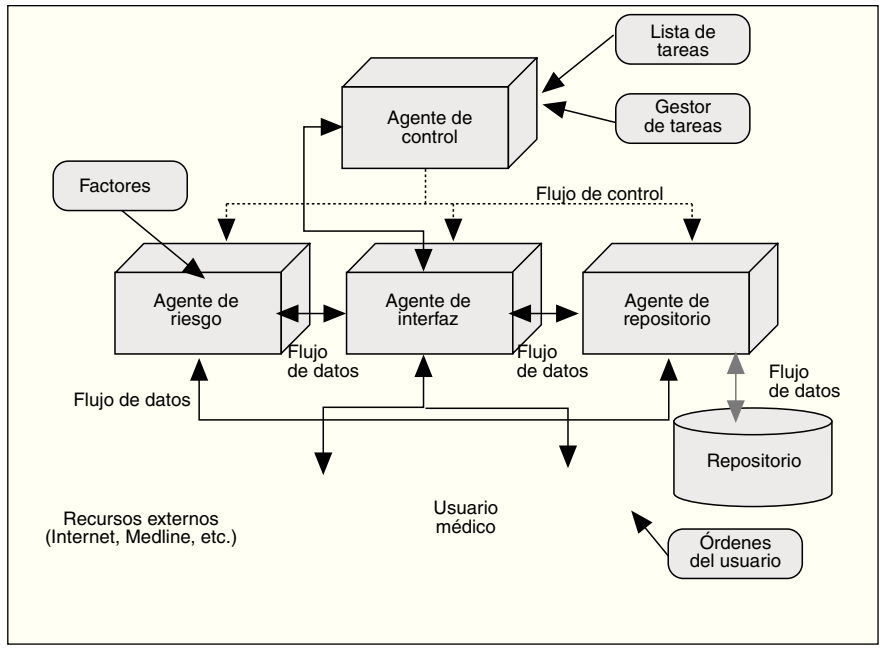

Figura 1. Modelo funcional preliminar del sistema completo.

que el segundo proyecto, consistente en la obtención del modelo de diseño y de un prototipo funcional restringido al núcleo del sistema, está actualmente en curso y los resultados del mismo son todavía parciales.

Modelo conceptual

El modelo conceptual obtenido contiene varios subsistemas que representan las grandes funciones descritas anteriormente y que se han recogido en la figura 1. Cada uno de estos subsistemas puede ser un sistema multiagente (abreviadamente, MAS), dando así un grado de flexibilidad y potencia mayores en el proceso de implementación, ya que pueden tratarse de sistemas heterogéneos basados en la tecnología que mejor se acople a las funciones asumidas.

En la tabla 3 se describen detalladamente las funciones de los subsistemas representados en el modelo conceptual. La existencia de un subsistema o agente de control determina que la arquitectura del sistema sea jerárquica, reduciendo así la carga de comunicación entre los subsistemas agentes con respecto a operaciones de negociación o competencia. Por ejemplo, no es recomendable en este sistema que la interfaz, el agente de repositorio o el agente de riesgo se dediquen a negociar o a competir por los recursos del sistema. Entonces será el agente de control, interpretando las órdenes del usuario y añadiendo las suyas propias, quien reparta la asignación de recursos.

Dados los objetivos generales del sistema, el subsistema más importante en el modelo es el agente de riesgo, designado para llevar a cabo el cálculo del riesgo (CHDR; *cardio heart disease risk*) para que un individuo determinado pueda padecer una enfermedad cardiovascular. En este subsistema se incorpora la estructura y comportamiento de cada uno de los elementos que se sabe tienen influencia en la variación del CHDR. Cada uno de estos elementos está representado inicialmente por un agente cuya base de conocimiento recoge lo que se sabe de ellos, pero tratando de encontrar un modo en el que puedan comunicarse entre ellos de forma estándar por medio de una arquitectura MAS.

Así, puede surgir el comportamiento emergente que aporte la solución buscada. Estos elementos son factores genéticos (el genotipo determinado por los polimorfismos de cada uno de los genes relacionados con las ECV) y los factores ambientales (dieta, estilos de vida, tabaco, alcohol, etc.)[23]. Los factores pueden ser de 2 signos: los de signo positivo son los factores de riesgo, y los de negativo, los factores de protección[24]. Por tanto, los agentes que representan al factor deben actuar según la naturaleza de éste.

Además, se deberán poder introducir factores desconocidos o no específicos, o perturbaciones en el sistema, de forma que se pueda modelar la evolución que puede sufrir un individuo, sobre todo por la influencia cambiante de los factores ambientales. Se pretende encontrar en el comportamiento

Tabla 3. Subsistemas de A-Genes

Subsistema	Funciones
Agente de control	Gestionará la interacción entre los restantes subsistemas. Parte del control estará programado para abordar tareas rutinarias; otra parte del control estará modelado mediante reglas que consideren actuaciones alternativas ante distintas variantes y reacciones del entorno y el resto de sistemas. Otra parte del control se limitará a traducir y trasladar las órdenes emitidas por el usuario humano
Agente de interfaz	Actuará como intermediaria entre el usuario y el sistema y entre el sistema e Internet. Incorporará agentes de búsqueda en Internet para reponer y añadir información actualizada en el repositorio según las órdenes emitidas por el agente de control. Las fuentes pueden ser Medline, OVID, bases de datos de SNP, etc.
Agente de repositorio	Gestionará toda la información que se almacenará en la base de datos (referida a estudios, proyectos…). Uno de los tipos de información contenidos serán los resultados de estudios epidemiológicos sobre muestras amplias de la población. Otras informaciones se relacionan con los estudios específicos y los datos de los individuos implicados. También se ha previsto almacenar la información proporcionada por sistemas de micromatrices de ADN
Agente de cálculo de riesgo	Modelará la estructura y comportamiento de cada uno de los elementos que se sabe tienen influencia en la variación del CHDR, en función del grado de conocimiento que se tiene de ellos, pero tratando de encontrar un modo en el que puedan comunicarse entre ellos de forma estándar. De esta comunicación puede surgir el comportamiento emergente que aporte la solución buscada. Cada uno de los elementos mencionados representa un factor de riesgo y ejecuta una parte de las ecuaciones de Framingham[9], la que le corresponde según el factor representado. Los factores pueden ser de 2 signos: los de signo positivo, que son los factores de riesgo, y los de signo negativo, que son los factores de protección[24]. Además, se permite la introducción de factores, desconocidos o no específicos, o perturbaciones en el sistema, de forma que se modela la evolución que puede sufrir un individuo sobre todo por la influencia cambiante de los factores ambientales (fig. 2)

CHDR: *cardio-heart disease risk*.

emergente del sistema la representación de la interacción gen-ambiente que determina la aparición y la prevalencia de las ECV.

La estructura de este subsistema también estará jerarquizada debido a que es necesario diseñar unos agentes dedicados a la planificación o tareas de nivel medio y otros a las de bajo nivel. Cada agente de nivel medio forma una *célula de riesgo* con un grupo de agentes de bajo nivel. Las tareas de nivel bajo son fundamentalmente los cálculos con los valores co-

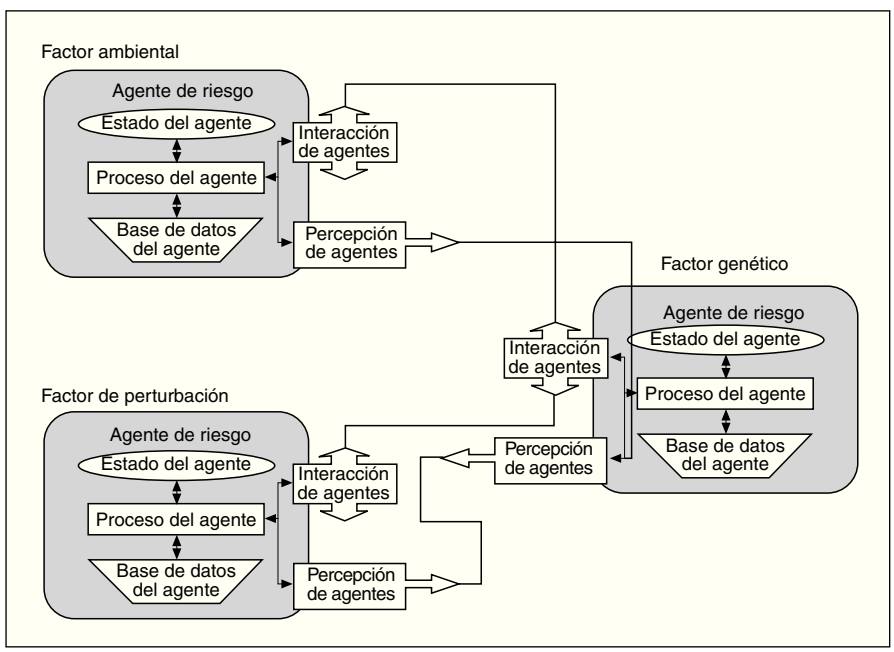

Figura 2. Estructura de interacción entre los agentes del subsistema de riesgo. Célula de riesgo.

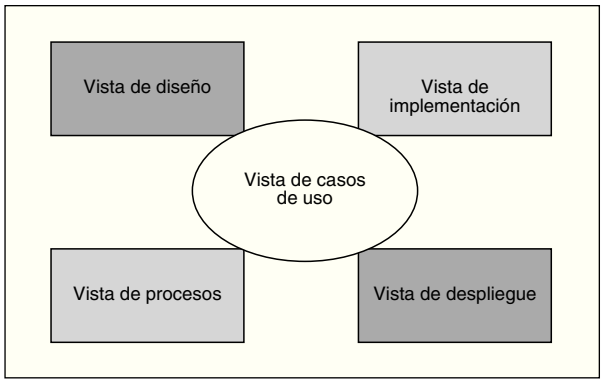

Figura 3. Modelo conceptual UML *(unified modelling language)*.

rrespondientes a cada uno de los factores de riesgo, y las de nivel medio son la síntesis de los cálculos mencionados antes en cada una de las ecuaciones de Framingham[9] y la comunicación con otras células de riesgo. Esto también facilitará el manejo de los datos proporcionados por sistemas de *microarrays* de ADN (fig. 2).

Por otra parte, considerando que gen, genotipo, fenotipo, etc. son entidades objetuales a modelar, se ha construido un modelo orientado a objetos para la gestión de la información genómica y cardiovascular usando como lenguaje de presentación y modelado UML. En este modelo se consideran tanto los elementos de información y las relaciones entre ellos (visión estructural) como su comportamiento en el sistema en funcionamiento (visión dinámica) (fig. 3).

El modelo conceptual obtenido está representado con UML y, por tanto, es un modelo basado, en principio, en un enfoque objetual. Sin embargo, dado que se han incorporado extensiones de UML para representar agentes y sistemas de agentes, el modelo se convierte en uno basado en un enfoque orientado a agentes. En la figura 4 se muestra parte del diagrama de clases del modelo de diseño (proceso unificado) que representa el subsistema de riesgo.

Este diagrama representa el proceso que incorpora el sistema para calcular el CHDR de un individuo en particular, asociado a un determinado estudio de población. Las clases que lo componen tienen las siguientes responsabilidades:

– *Clase agente de riesgo*: representa al agente encargado de realizar el CHDR para un individuo concreto.

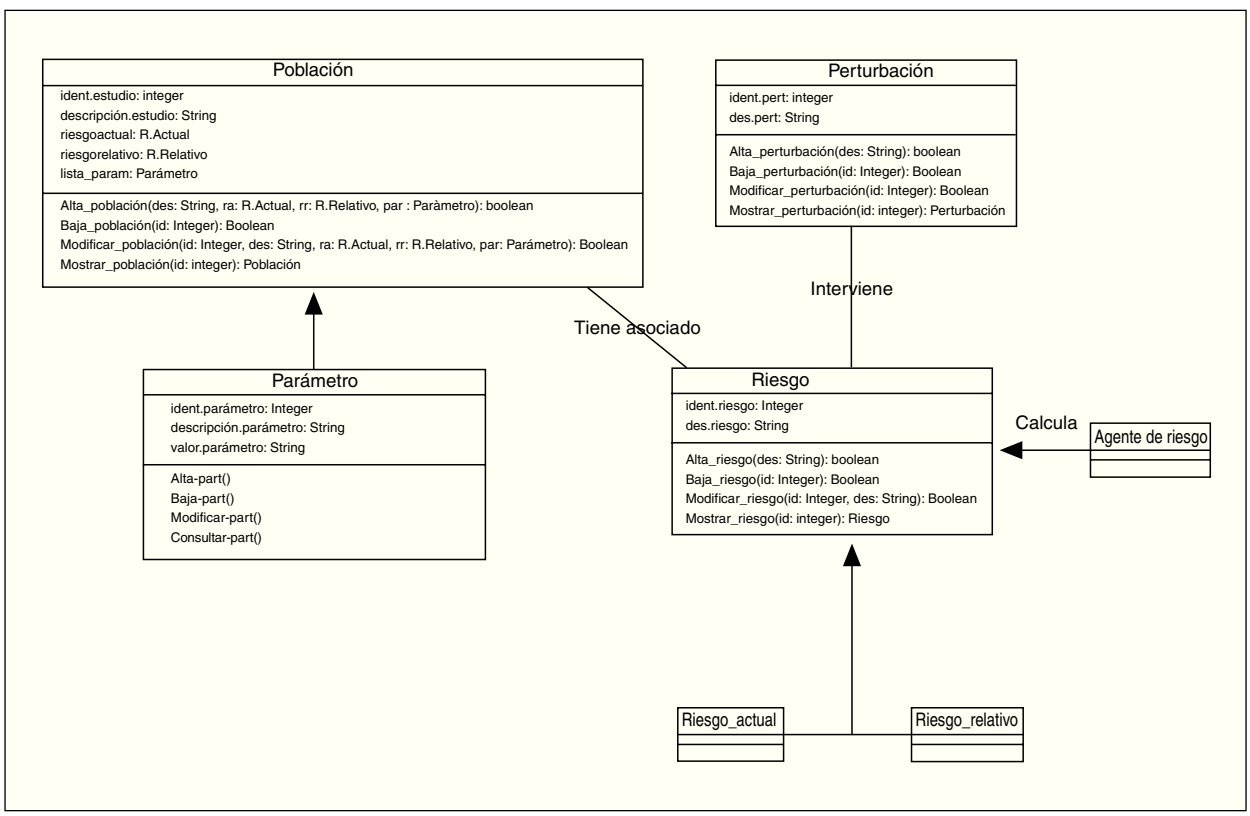

Figura 4. Diagrama de clases del cálculo del riesgo (subconjunto).

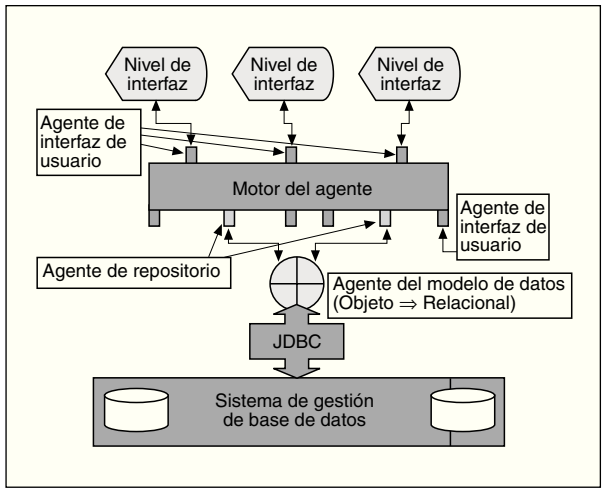

Figura 5. Arquitectura de *software* del sistema.

– *Clase riesgo*: representa el riesgo cardiovascular calculado a partir de los datos de un individuo perteneciente a una muestra de población y que es realizado por el agente anterior. Esta clase es una generalización y de ella se especializan otras 2 clases que son riesgo relativo y riesgo actual.

– *Clases riesgo actual y riesgo relativo*: son una especialización de la clase riesgo. El riesgo actual representa la probabilidad pronosticada de que un paciente experimente un episodio cardíaco en los próximos 10 años (tiempo fijado según el Framingham Heart Study)[9]. El riesgo relativo es el riesgo cardiovascular basado en los factores de riesgo individuales.

– *Clase perturbación*: representa cualquier alteración introducida en el sistema, ya sea un factor genético o ambiental no identificado como una perturbación general que afecta al comportamiento global del sistema.

– *Clase población*: representa el conjunto de resultados de estudios epidemiológicos y clínicos sobre poblaciones que se almacenan en el sistema, como la fuente de información a partir de la que se extraen los datos del individuo del que se calculará el riesgo cardiovascular.

Diseño

La naturaleza heterogénea de los entornos utilizados apoya la adopción de una arquitectura de agentes. Por tanto, se ha diseñado la arquitectura de *software* de forma que se distribuye en 3 capas: interfaz, motor de agente y base de datos (fig. 5).

El nivel de interfaz está desarrollándose en Delphi, ya que permite mejor flexibilidad y rapidez para obtener una interfaz de usuario lo más potente posible. El motor de agente está desarrollándose con Agentbuilder y dará lugar a módulos Java que representan los distintos agentes del sistema. La utilización de Java permite que se utilice una base de datos independiente del resto de elementos del sistema y, mediante el protocolo JDBC (Java Data Base Connectivity), puede utilizarse, por ejemplo, cualquier gestor estándar como Oracle, SQL Server, DB2, etc. Si el modelo de datos del gestor es objeto-relacional, la conexión con el resto de la arquitectura es casi directa. Sin embargo, si se trata

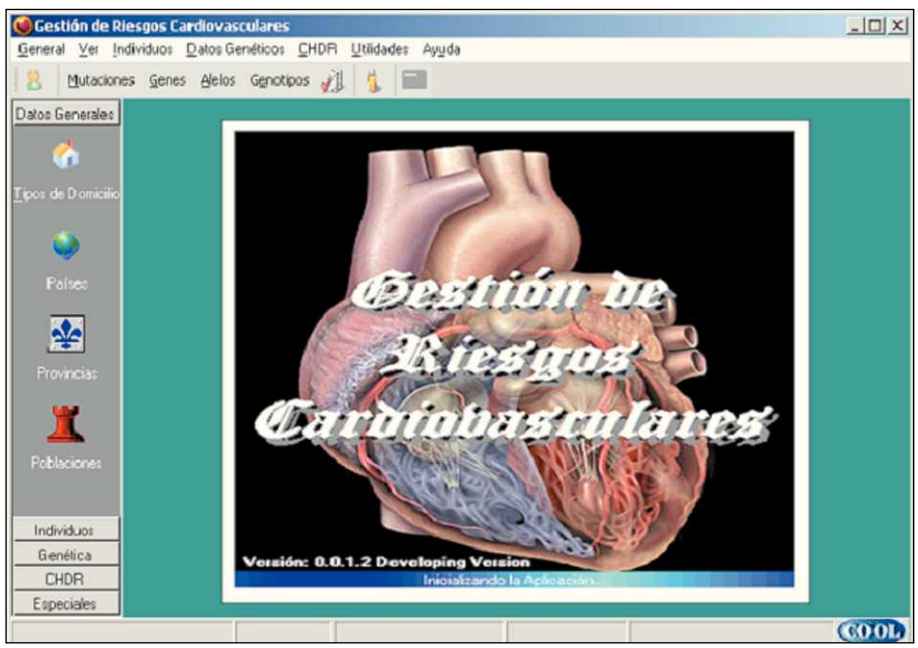

Figura 6. Arquitectura de *software* del sistema.

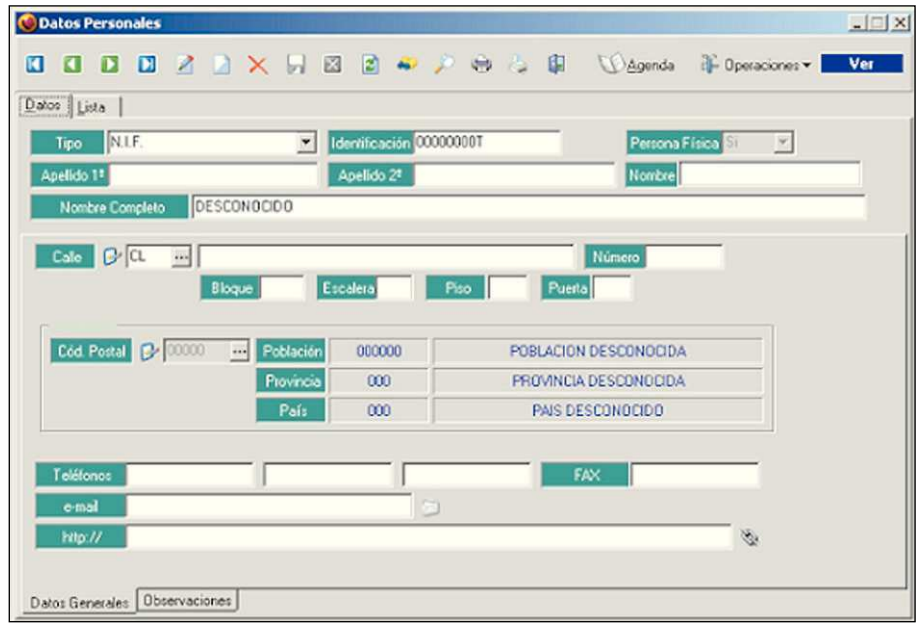

Figura 7. Pantalla de petición de información relativa a un individuo de estudio.

de un modelo relacional es necesario introducir un componente adicional que haga de interfaz entre el sistema y la base de datos.

Prototipo

Se ha desarrollado un prototipo funcional que cubre las operaciones de control general y de adquisición de datos de cada individuo. El sistema se caracteriza por su facilidad para poderse adaptar y cambiar de modelo o de características individuales a tener en cuenta, sin necesidad de modificar el programa. También cabe destacar la posibilidad de presentar su interfaz en diferentes idiomas, pudiendo el usuario extenderlos sin necesidad de intervención por parte del programador. Estas características también se han podido incorporar a este prototipo.

El prototipo consta de un conjunto de pantallas, empezando por la que ejerce de portal de la aplicación hasta las pantallas de introducción de información acerca de genes, polimorfismos, mutaciones, etc. (fig. 6).

En la parte superior de cada pantalla aparece una barra de navegación, que permite ir a registros anteriores o posteriores, introducir nuevos registros, modificar registros ya introducidos o borrar algunos de estos registros. También se incluyen botones con funciones de búsqueda o localización de un registro concreto o de un determinado conjunto de registros y de generación de informes sobre los registros introducidos. En cada función aparecen 2 tipos de botones, a cada lado del lugar designado, para la introducción de los datos solicitados. El botón de la izquierda permite acceder de forma rápida a la pantalla donde se mantiene la información del dato solicitado, mientras que el de la derecha permite realizar una búsqueda sobre los elementos ya introducidos previamente.

En algunas pantallas aparecen botones adicionales para realizar más operaciones que las detalladas anteriormente, pero estas opciones dependen de cada pantalla concreta y de la función que tienen asignada. Así, por ejemplo, en el caso de la pantalla de petición de información relativa a un individuo de estudio, en el campo reservado al identificador de la persona se puede acceder tanto a la pantalla donde se introducen las personas para darlo de alta o modificar sus datos, como a la lista de búsqueda para seleccionar la persona deseada (fig. 7).

En la figura 7 se expone la forma en que se introduce la información personal de cada uno de los individuos de estudio, que permite la confección de cartas dirigidas a los individuos de estudio para comunicarles, por ejemplo, resultados del cálculo de su riesgo cardiovascular absoluto y relativo. Dadas las características con las que se ha desarrollado el prototipo, también se permite introducir organismos, tales como institutos de investigación o universidades, que puedan estar realizando determinados proyectos de investigación, para llevar un control sobre éstos. Por último, esta pantalla también se puede utilizar a modo de agenda o listín telefónico.

En la actualidad, se están desarrollando las pantallas que permiten la introducción de los modelos

de cálculo del riesgo cardiovascular y de realización de dicho cálculo.

Discusión

El CHDR individual se puede calcular con precisión mediante una aproximación multiagente que puede manejar un amplio conjunto de factores genéticos y ambientales en un espacio de complejidad polinomial. Es necesario explotar las características de los sistemas multiagentes heterogéneos y diseñar los elementos necesarios para obtener un comportamiento emergente que permita manejar la complejidad de interacciones entre los distintos factores de riesgo.

El presente trabajo presenta un modelo conceptual, con UML, de un sistema basado en una arquitectura MAS para el cálculo individual del riesgo cardiovascular individual como soporte de estudios médicos y epidemiológicos. El trabajo ha intentado tomar las aproximaciones teóricas, metodológicas y prácticas de las disciplinas inteligencia artificial e ingeniería del *software*, que pueden trabajar juntas en el marco de la bioinformática aplicada a la faceta de la epidemiología genómica de las enfermedades cardiovasculares. Dado que se trata solamente de un modelo conceptual, se está estudiando en la actualidad la aplicación y adecuación de distintos enfoques de razonamiento para implementar los agentes individuales y también la estructura multiagente. En la actualidad, ya se dispone de un prototipo funcional general que cubre algunas de las funciones de interfaz del sistema e incorpora requisitos de diseño para mejorar la flexibilidad y el manejo del sistema por los usuarios.

Este trabajo ha presentado algunos de los resultados de 2 de los primeros subproyectos de un proyecto de gran envergadura técnica y temporal denominado A-Genes, realizado con la colaboración de investigadores de Estados Unidos y España, como una herramienta que permita en un futuro llegar a calcular el riesgo individual con precisión y como una operación más de un acto médico.

Agradecimientos

Las primeras fases del trabajo de investigación, diseño y desarrollo de A-Genes se realizaron cuando Oscar Coltell y Dolores Corella estuvieron realizando una estancia como profesores invitados en el Nutrition and Genomics Laboratory del Human Nutrition Research Center on Aging at Tufts University, Boston (Estados Unidos), durante los veranos de 2001 y 2002. Posteriormente, en 2003, este proyecto ha sido financiado parcialmente por la red temática "G03/160. INBIOMED. Plataforma de almacenamiento, integración y análisis de datos clínicos, genéticos, epidemiológicos e imágenes orientada a la investigación sobre patologías", del Instituode Salud Carlos III.

Bibliografía

1. Bishop M, editor. Guide to human genome computing. 2nd ed. London: Academic Press, 1998.
2. Nakamura Y. Human genome analysis and medicine in the 21st century. Proceedings of the fourth annual international conference on computational molecular biology 2000, Tokyo, Japan. New York: ACM Press, 2000; p. 221-2.
3. Bader GD, Hogue CWV. BIND–A data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 2000;16:465-77.
4. Fischer C, Schweigert S, Spreckelsen C, Vogel F. Programs, databases, and expert systems for human geneticists-a survey. Hum Genet 1997;97:129-37.
5. Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment Interaction: an overview of available methods. Epidemiol Rev 1998;20:137-47.
6. Bridge PJ. The calculation of genetic risks. Philadelphia: Johns Hopkins University Press, 1997.
7. Isles C, Ritchie L, Murchie P. Risk assessment in primary prevention of coronary heart disease: randomized comparison of three scoring methods. BMJ 2000;320:690-1.
8. Wallis E, Ramsay L, Ul Haq I. Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 Scottish health survey population. BMJ 2000;320:671-6.
9. Wilson PWF, D'Agostino RB, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.
10. Baker S, Priest P, Jackson R. Using thresholds based on risk of cardiovascular disease to target treatment for hypertension: modeling events averted and number treated. BMJ 2000;320:680-5.
11. Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES. Class prediction and discovery using gene expression data. Proceedings of the fourth annual international conference on Computational molecular biology 2000, Tokyo, Japan. New York: ACM Press, 2000; p. 263-72.
12. Barash Y, Friedman N. Context-specific Bayesian clustering for gene expression data. Proceedings of the fifth annual international conference on Computational biology 2001. Montreal, Quebec, Canada. New York: ACM Press, 2001; p. 12-21.
13. Southern E. DNA microarrays–The how and the why. Proceedings of the third annual international conference on computational molecular biology 1999. New York: ACM Press NY, 1999; p. 311-2.
14. Filkov V, Skiena S, Zhi J. Analysis techniques for microarray time-series data. Proceedings of the fifth annual international conference on Computational biology 2001, Montreal, Quebec, Canada. New York: ACM Press, 2001; p. 124-31.
15. Weiss G, editor. Multiagent systems. Modern approach to distributed artificial intelligence. Cambridge: MIT Press, 1999.
16. Elammari M, Lalonde W. An agent-oriented methodology: high-level and intermediate models. Proceedings of the Agent-Oriented Information Systems conference (AOIS) 1999. Heidelberg, Germany. Disponible en: http://www.carleton.ca/~elammari/AOIS99/
17. Ordovás JM, Corella D, Cupples LA, Demissie S, Kelleher A, Coltell O, et al. Polyunsaturated fatty acids modulate the effects of the APOA1 G-A polymorphism on HDL-cholesterol concentrations in a sex-specific manner: the Framingham Study. Am J Clin Nutr 2002;75:38-46.
18. Jacobson I, Booch G, Rumbaugh J. El proceso unificado de desarrollo de software. Madrid: Addison-Wesley, 2000.
19. Rumbaugh J, Jacobson I, Booch G. El lenguaje unificado de modelado. manual de referencia. Madrid: Addison-Wesley, 2000.
20. SPSS genlog1, 2002 [consultado 18/09/2002]. Disponible en: http://www.spss.com/tech/stat/algorithms/11.0/genlog_multinomial.pdf
21. SPSS genlog2, 2002 [consultado 18/09/2002]. Disponible en: http://www.spss.com/tech/stat/algorithms/11.0/genlog_poisson.pdf
22. Niyonsenga T, Xhignesse M, Courteau J, Ciampi A, Lussier-Cacan S, Roy M. Desarrollo de una escala de riesgo para evaluar el riesgo actual de cardiopatía coronaria empleando variables de la historia familiar. Cardiovascular Risk Factors 2000;9:30-42.
23. Cambien F, Tiret L. Genotipo y riesgo de la coronopatía. Cardiovascular Risk Factors 1998;7:93-101.
24. Sentí M, Ordovás JM. Comprender los efectos de las interacciones protectoras entre los factores genéticos y ambientales sobre las enfermedades isquémicas cardiovasculares. Cardiovascular Risk Factors 1999;8:46-54.