ENKI-DB: sistema de información taxonómica y molecular de especies propias de la biodiversidad colombiana

ENKI-DB: Molecular and taxonomic data integration system for Colombian species

Andrés M. Pinzón[**], *María Teresa Reguero*[***], *y Emiliano Barreto*[*]

RESUMEN

ENKI-DB, es un sistema de información taxonómica y molecular de especies propias de la biodiversidad colombiana, accesible vía Internet. Mediante ENKI-DB se tiene acceso inmediato a toda la información taxonómica y molecular presente en las bases de datos SPICA®, SIB, EMBL y UNIPROT para especies propias de la biodiversidad colombiana exclusivamente. Hasta la fecha el sistema ha logrado enlazar 10 808 registros de especies propias de la biodiversidad colombiana (presentes en las bases de datos SIB y SPICA®), para las que han encontrado 1 976 751 registros moleculares, 96 337 provenientes de las bases de datos de información proteica (UniprotKb, Uniref y Uniparc) y 1 880 414 de la base de datos EMBL de DNA. Aunque esta información se encuentra presente también, de manera independiente, en cada una de estas bases de datos, a través de ENKI- DB los datos son accesibles desde una sola interfase y de manera integrada y depurada. El sistema permite además realizar alineamientos locales utilizando la implementación del algoritmo BLAST del NCBI.

ENKI- DB ha sido desarrollado en su totalidad en PHP y PERL, haciendo uso de las librerías Bio-PHP, Bio-PERL y utilizando la base de datos MySQL como repositorio central de información. ENKI-DB es accesible vía Internet, de manera completamente gratuita y pública en la siguiente dirección: http://bioinf.ibun.unal.edu.co/enkidb/
Contacto: cenbio_nal@unal.edu.co

Palabras clave: ENKI, molecular, taxonomía, biodiversidad colombiana, bioinformática.

ABSTRACT

ENKI-DB is a taxonomic and molecular integration system for Colombian species which is available on the internet. This system provides users with immediate access to all taxonomic and molecular data present in SPICA®, SIB, EMBL and UNIPROT databases for all Colombian species. To date, our system has been able to link 10,808 Colombian species (SIB and SPICA® databases) to 1 976 751 molecular entries, 96 337 from UniprotKb, Uniref and Uniparc and 1,880,414 from the EMBL DNA database. Although this information is also available on each database independently, ENKI-DB allows users to reach all that integrated information through one interface. We have also integrated the NCBI-BLAST programme with the ENKI-DB system which allows users to run pair-wise comparisons on the same ENKI-DB web interface. ENKI-DB has been developed with Bio-PHP and Bio-PERL and uses MySQL RDBMS as backup.

Availability: http://bioinf.ibun.unal.edu.co/enkidb/
Contact: cenbio_nal@unal.edu.co

Key words: ENKI, molecular integration system, taxonomy, Colombian biodiversity, bioinformatics.

Recibido: mayo 05 de 2006 Aceptado: junio de 2006

* MSc en Farmacología. Centro de Bioinformática, Instituto de Biotecnología, Universidad Nacional de Colombia. Correo electrónico: ebarretoh@unal.edu.co

** Biólogo. Centro de Bioinformática, Instituto de Biotecnología, Universidad Nacional de Colombia. ampinzonv@unal.edu.co

*** MSc en Ciencias Químicas. Centro de Bioinformática, Instituto de Biotecnología, Universidad Nacional de Colombia. Correo electrónico: mtregueror@unal.edu.co

INTRODUCCIÓN

Durante décadas, los repositorios digitales y públicos de información biológica han sido parte integral y crítica de la investigación científica alrededor del mundo. Gracias a este tipo de sistemas es posible no sólo centralizar la información biológica, producto de esfuerzos simultáneos alrededor del mundo por diversos proyectos de investigación, sino que además pueden estar disponibles inmediatamente, para cualquier persona en el mundo con acceso a Internet.

En este sentido existen innumerables esfuerzos que han cambiado drásticamente la manera de entender y acceder a la información biológica, en la actualidad existen 858 bases de datos que contienen principalmente información biológica molecular (Galperin, 2006) y que cubren casi todos los rangos posibles de esta: rutas metabólicas (Kanehisa, et ál., 2006); DNA (Cochrane, et ál., 2006); proteínas (Apweiler, et ál., 2004; Berman, et ál., 2003), etc. Se destacan dentro de estas bases de datos el European Bioinformatics Institute (EMBL, 2006) como una colección de secuencias de ADN y ARN que diariamente es actualizada y sincronizada con las bases de datos National Center for Biotechnology Information (GENBANK, 2006) y National Institute of Genetics of Japan (DDBJ, 2006), generando igualdad de contenido en las tres y que actualmente cuenta con más de 60 millones de registros y UNIPROT (Uniprot, 2006) con más de 2,3 millones de registros de secuencias de proteínas y su información funcional, la cual incluye información de la tres más importantes bases de datos de proteínas: SwissProt, TrEMBL (Expasy Proteomics Server, 2006) y PIR (Georgetown University Medical Center, 2006).

Además de la facilidad de acceso y de la constante actualización de estos sistemas de información, existe una característica relevante común a todos ellos: la capacidad de crear nexos entre su información y otras bases de datos, llegando a crear complejas redes de información biológica, en las cuales cada fuente puede conectar directa o indirectamente con el resto de información asociada al recurso que ofrece. Es esta característica la que permite que a partir de una simple búsqueda de por ejemplo una secuencia proteica en SwissProt se pueda acceder no sólo a dicha secuencia, sino al gen que la origina en la base de datos EMBL; a su posible relación con desórdenes genéticos en la base de datos OMIM y a su estructura terciaria en la base de datos PDB, por nombrar solamente algunas posibilidades.

De esta manera se puede constatar que, si bien la sistematización de la información biológica ha sido un gran avance para el conocimiento del mundo natural, el hecho que nos lleva cada vez más cerca de su comprensión es la interconexión entre las diversas fuentes y tipos de datos, lo que permite, aproximarse de una manera más real, a la comprensión de que los organismos no están aislados de su entorno, ni de sus estructuras moleculares o funciones biológicas, de modo que es posible encontrar nuevas relaciones entre todos estos elementos, relaciones que a primera vista no son evidentes.

Por otra parte, a pesar de la gran cantidad de información biológica disponible actualmente, la inmensa mayoría de los datos interconectados y de dominio público son de carácter molecular y la información taxonómica, que ha sido sistematizada desde hace mucho tiempo se encontraba, en su mayoría, aislada de los datos moleculares. Este hecho tiene sus raíces en la manera como se han desarrollado las áreas científicas alrededor de esta información. En general los datos taxonómicos se han mantenido reservados tal vez por el cuidado que los investigadores tienen con las colecciones o por la oportunidad de generar alguna publicación científica. En el caso de los datos moleculares en general se han mantenido en el dominio público y con la tendencia a compartir las herramientas informáticas, generando proyectos interinstitucionales e internacionales, como es el caso del proyecto del genoma humano.

En cuanto a los datos taxonómicos, desde hace unos diez años se inició un proceso similar al acaecido en el caso de los datos moleculares. Es así como se han diseñado proyectos internacionales que buscan superar la escasa interconexión y comunicación entre grupos de investigación y eliminar la redundancia entre las diferentes colecciones de datos biológicos. Esto permitiría superar el hecho de que sólo el 2% de las especies conocidas a nivel mundial tienen sistematizados los

datos morfológicos y a que dado el desarrollo principalmente de la biología molecular casi hayan desaparecido los taxónomos expertos de los que depende en gran medida la calidad de los datos. Por tanto, en una clara extensión de la bioinformática al campo de la biodiversidad se ha buscado hacer accesible, de manera digital, la enorme cantidad de información de la biodiversidad global, haciendo que dicha información sea consistente y compatible entre diversos sistemas informáticos a través de la implementación de iniciativas globales como: Species 2000 (Species 2000, 2006) una federación de organizaciones de bases de datos que trabaja en conjunto con usuarios, taxónomos, y agencias patrocinadoras como United Nations Environment Programme UNEP y la Global Environment Facility GEF, buscando incluir dentro de sus registros a todos los organismos conocidos en la tierra (40% del total de especies conocidas hasta ahora); o el Global Biodiversity Information Facility (GBIF, 2006) cuya misión es similar a la de Species 2000, apoyada por la Organización para el Desarrollo y Cooperación Económica (OCDE) ó el proyecto árbol de la vida (Tree of Life web Project, 2005) que intenta proveer información acerca de diversidad, historia evolutiva y características de los organismos en la tierra con la colaboración de biólogos alrededor del mundo.

Por otro lado, en coordinación con las iniciativas globales se desarrollan proyectos en los diferentes continentes como The Biological Collection Access Service (BioCASE, 2006), que busca brindar un servicio de acceso a información de las colecciones biológicas en Europa, vinculando más de 30 instituciones en su mayoría como nodos nacionales que se encargan del mantenimiento de la información que es compartida a través de un sistema que intercambia metadatos y que permite consultar cualquiera de las bases de datos como si fueran una sola. También en América se encuentran iniciativas similares como el Sistema de Información Integrado Taxonómico de Norteamérica, (ITIS, 2006) base de datos de fácil acceso, con información confiable sobre las especies de México, Estados Unidos de América y Canadá y su clasificación jerárquica; Comisión Nacional para el Conocimiento y Uso de la Biodiversidad en México (CONABIO) con imágenes satelitales en línea, información de colecciones, taxónomos y especies nativas mexicanas; y el Centro de Investigación y Gestión de la Biodiversidad en Costa Rica (INBio, 2006) establecido para apoyar los esfuerzos para conocer la diversidad biológica y promover su uso sostenible en Costa Rica.

En Colombia también existen varias iniciativas para la recopilación de información biológica, dentro de las cuales se pueden mencionar, el Sistema de Información sobre Biodiversidad (SIB, 2006) del Instituto Alexander von Humboldt y el Sistema de Información Biótico Ambiental (SPICA®, 2006) del Instituto de Ciencias Naturales de la Universidad Nacional de Colombia, entre otras. El Sistema de Información sobre Biodiversidad es una iniciativa de carácter nacional, encaminada a satisfacer las necesidades de información del país en cuanto a la conservación y el uso sostenible de sus recursos biológicos. El Instituto von Humboldt es la entidad responsable de la coordinación y puesta en marcha de dicho sistema que involucra otras entidades como universidades y centros de investigación, de acuerdo con su papel en la gestión y generación de datos e información sobre biodiversidad. El proceso de implementación del SIB gira en torno a tres ejes principales: capacidad para gestionar eficientemente datos e información sobre biodiversidad entre las diferentas instituciones, infraestructura que incluye los elementos físicos (*hardware*), lógicos (*software*), los estándares y la arquitectura del sistema, entre otros, y contenido de información con el que se promueve la generación y disponibilidad de datos e información interoperables, coherentes y pertinentes a los diferentes fines definidos para el sistema. En la actualidad el sistema ha avanzando en la integración de los archivos de autoridad taxonómica y en la integración de las bases de datos de diferentes actores entre los cuales se incluyen centros de investigación y universidades, a través de una estructura de intercambio de metadatos que sigue estándares internacionales, lo que facilita la interoperabilidad del sistema.

El Sistema Biótico Ambiental —SPICA®— registrado, desarrollado e implementado por la Universidad Nacional de Colombia, es un sistema de información sobre flora y fauna colombiana, diseñado para la articulación de diversos sectores de

investigación de las ciencias naturales, sociales y económicas. Fue desarrollado para el manejo de colecciones biológicas de: entomología, ornitología, herpetología, mastozoología e ictiología, y planteado como un sistema de proyección muy amplia que puede ser adecuado a las diferentes necesidades del estudio de la biota. De hecho su implementación en diversas instituciones ha permitido, entre otras, la construcción del Sistema de Información del Chocó biogeográfico, realizado por el Instituto de Investigación Ambiental del Pacífico —IIAP— con apoyo de la Universidad Nacional de Colombia; del Sistema de Estadísticas Forestales de Colombia para la caracterización y divulgación de la flora forestal, existiendo la posibilidad de consulta directa sobre aproximadamente 945 especies y de una plataforma de trabajo de varios proyectos del Instituto de Hidrología, Meteorología y Estudios Ambientales —IDEAM—.

De forma general, tanto para las colecciones de datos moleculares como para las de biodiversidad, no existe un mecanismo de conexión directa de unas con otras, es decir no se encuentra una manera en la que, a través de una base de datos molecular, se pueda recuperar la información morfológica y taxonómica del organismo que dio origen a la secuencia, ubicándolo dentro de la biodiversidad de un determinado lugar y viceversa, pues no existe un nexo claro, explícito que las una y los investigadores se encuentran frecuentemente con la frustrante experiencia de tener que revisar diversos sistemas de datos para obtener la misma información que deberíamos obtener fácilmente a través de uno sólo, y sin información acerca de la procedencia del organismo a partir del cual fue obtenida la secuencia. Este problema también se observa en el entorno colombiano donde la información sobre secuencias de ácidos nucleicos y proteínas, de organismos nativos de Colombia, se encuentra dispersa, ya que cada grupo genera y almacena sus propios datos y una gran parte de ellos se encuentran en los bancos de datos mundiales sin referencia a su origen y aunque está previsto su almacenamiento en sistemas como SPICA® y SIB, aún éstos no cuentan con este tipo de datos.

Es claro por tanto, que es necesaria la interconexión entre los sistemas de información de biodiversidad y bases de datos de secuencias ácidos nucleicos y proteínas como el EMBL y Unipro, que mantienen la mayor colección de datos moleculares del mundo, a través de un sistema que realice los nexos necesarios entre la información taxonómica de las especies y su información molecular asociada.

En esta dirección este trabajo presenta a ENKI-DB como prototipo de un sistema de información diseñado para el establecimiento de dichos nexos moleculares y taxonómicos, agregando además un componente de vital importancia para la información sobre biodiversidad que es mantener la relación entre una secuencia y el organismo y su procedencia, en este caso dedicada a organismos exclusivamente presentes en Colombia.

Sistema de integración de datos ENKI. El sistema ENKI-DB se orienta a la creación de un repositorio de datos moleculares relacionados con organismos propios de la biodiversidad colombiana, con la capacidad de interactuar con sistemas que almacenan otro tipo de información biológica (hábitat nativo, taxomonia, etc.) almacenados en sistemas de información en Colombia, como el SIB y SPICA®,como se repre-

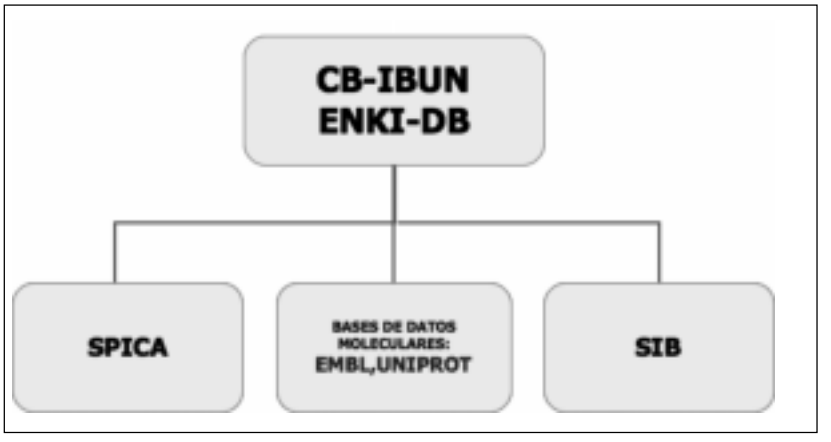

Figura 1. Modelo sistema ENKI-DB.

senta en al figura 1. Las fuentes de información molecular que alimenta a ENKI-DB son las bases de datos EMBL y UNIPROT de secuencias de ácidos nucleicos y proteínas respectivamente, ya que son las más grandes y completas a nivel mundial.

El modelo ENKI-DB (figura 1) se estructuró con la idea de que interactuara con cada una de las bases de datos tanto moleculares (EMBL, UNIPROT) como de biodiversidad (SIB, SPICA®), de tal forma que se pudieran cruzar sus datos sin tener la necesidad de tener las colecciones de datos completos de manera local, de cada una de ellas y así obviar la enorme capacidad de almacenamiento que se requeriría y más importante aun, evitar los problemas técnicos que implica el desarrollo de un sistema que unifique las diversas tecnologías que fueron empleadas para la implementación de estas bases de datos, que van desde una implementación en Oracle en el caso de SPICA® hasta archivos en texto plano (EMBL, UNIPROT), pasando por manejo de metadatos y formato XML (SIB).

Al evaluar las diferentes bases de datos y las necesidades de los posibles usuarios del sistema ENKI – DB (figura 2) se estableció que el nombre de la especie y la secuencia (de nucleótidos o aminoácidos) serían los datos requeridos por un usuario que quisiera consultar el sistema y obtener el cruce entre el nombre de la especie y los datos moleculares asociados a ella, siempre que se trate de una especie reportada como parte de la biodiversidad colombiana en las bases de datos SIB o SPICA®. Tambien, si el usuario intenta ver si hay organismos que tengan secuencias parecidas a una secuencia problema, reportados dentro de la diversidad colombiana, debería conseguir dicha información a través de una búsqueda BLAST, (McGinnins, et ál., 2004), implementada específicamente dentro de ENKI- DB y recuperar rápidamente los registros correspondientes a secuencias similares que estén incluidas en EMBL y UNIPROT y cuyas especies estén reportadas en el SIB o en SPICA®.

El diagrama de casos de uso (figura 2) muestra que el sistema ENKI- DB interactúa con las bases

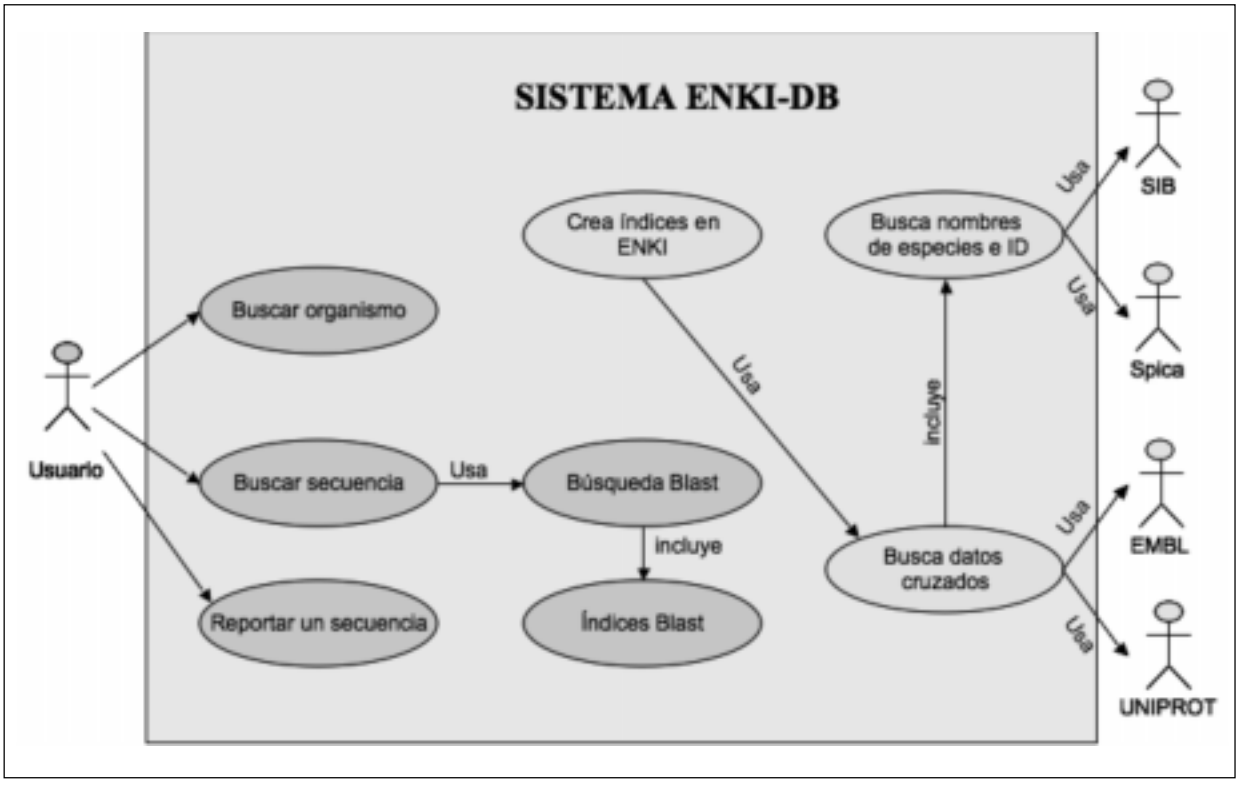

Figura 2. Diagrama de casos de uso sistemas ENKI-DB

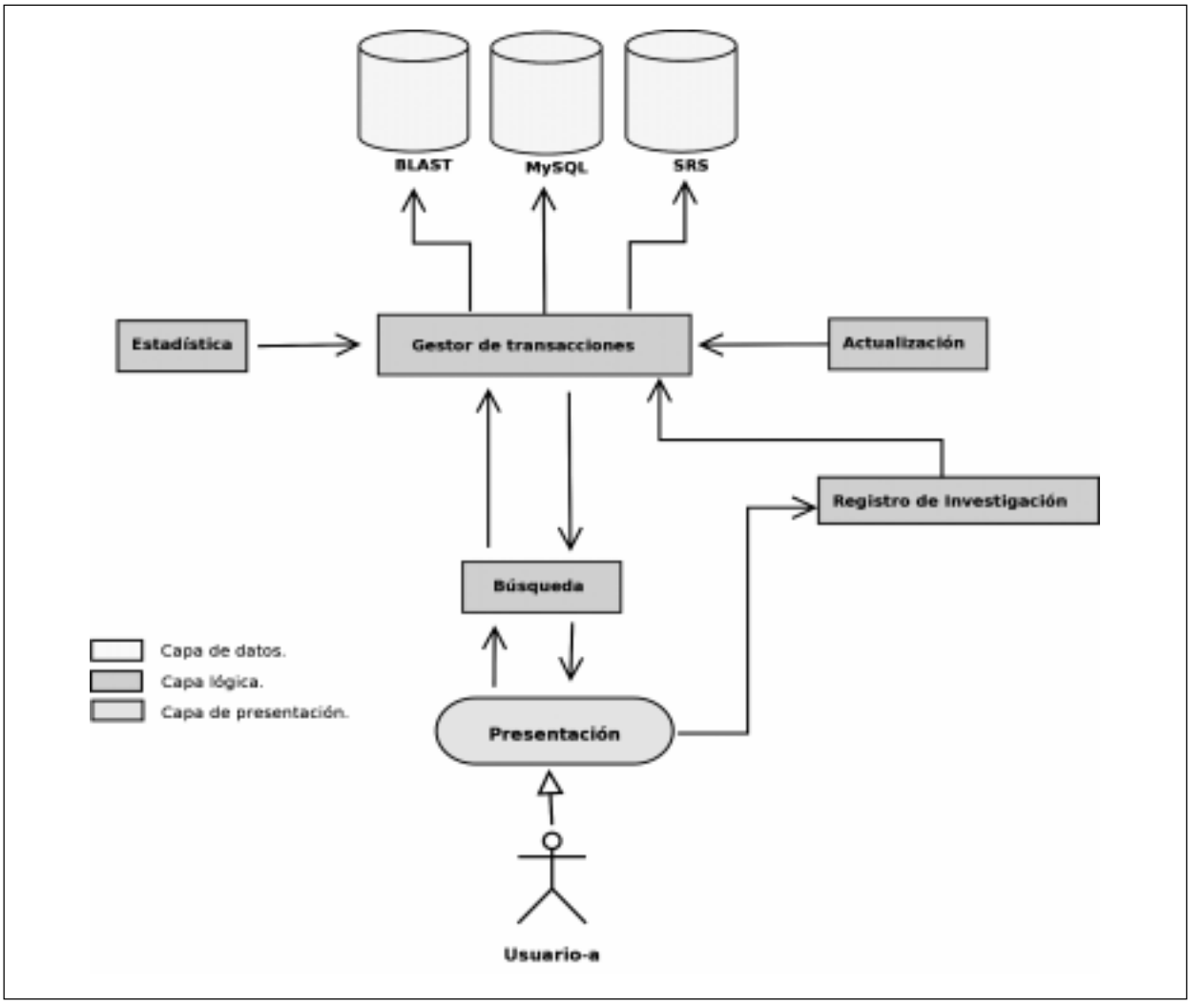

Figura 3. Modelo de integración de datos del sistema ENKI-DB

de datos SIB y Spica®, recuperando un listado con los nombres de las especies almacenadas en cada una de ellas y un código identificador (ID) que sirve de referencia dentro de cada una para recuperar rápidamente la información contenida en la misma. Con la lista de nombres de especies se procede a hacer una búsqueda de las secuencias reportadas para cada especie en las bases de datos EMBL y UNIPROT, para lo cual se utiliza el servidor del Sistema de Recuperación de Secuencias (SRS), implementado para tal fin, (http://srs.ibun.unal.edu. co:8080/srs81/) y que se encarga de hacer esta búsqueda especializada con la que se obtienen los datos y referencias que alimentan a ENKI y que permiten crear los índices cruzados entre las diferentes bases datos.

El sistema de integración de datos de origen taxonómico y molecular ENKI-DB, se articuló en 3 grandes capas (figura 3):

La primera capa denominada capa de datos corresponde al almacenamiento físico de datos, el cual es realizado utilizando MySQL versión 4.0.18, el Sistema de Recuperación de Secuencias (SRS), bases de datos especialmente formateadas que operan conjuntamente con el programa de análisis de secuencias Basic Local Aligment Search Tool (BLAST) y el sistema de archivos propio del sistema operativo SuSE Linux 9.0 de Linux, operando en un servidor SunV40z, AMD Opteron 1800Mhz (x2), 2Gb RAM.

La capa lógica está conformada por el conjunto de rutinas de programación que permiten la ejecución de las solicitudes de usuario y de las peticiones propias del sistema, desarrolladas en lenguaje PHP (http://www.php.net) versión 4.3.4, utilizando algunas subrutinas PERL (http://perl.org) versión 5.8.3 y los módulos BioPerl (http://www.bioperl.org) y BioPHP versión 1.0 (http://www.biophp.org). Esta capa opera todos los procesos concernientes a la actividad de ENKI-DB y como se puede ver en la figura 3 está conformada por los siguientes módulos:

- Búsqueda. Constituido por el conjunto de rutinas e interfaces que permiten al usuario realizar peticiones al sistema.

- Actualización. Encargado de mantener actualizados los registros tanto moleculares como taxonómicos del sistema, es el más complejo de todos y sobre él reposa la actividad principal de ENKI –DB, ya que sus cuatro submódulos se encargan de realizar los procesos de actualización semanal de los datos provenientes SPICA® y el SIB; la búsqueda y actualización de información molecular asociada a cada una de las especies presentes; el registro de dicha información en el sistema de bases de datos de ENKI-DB; y la creación de las entradas apropiadas en la base de datos BLAST-ENKI.

- Registro de investigación. Permite a investigadores nacionales e internacionales registrar directamente la información molecular derivada de sus investigaciones en biodiversidad colombiana en ENKI-DB.

- Estadística. Encargado de generar los datos estadísticos sobre el sistema, tales como el número total de secuencias, el número de especies presentes, tiempos de operación, fechas de actualización, uso de disco y en general, información relevante para el mantenimiento del sistema.

- Gestión de transacciones. es el encargado de gestionar las negociaciones con las bases de datos, como es de esperarse su papel es central en la operación de la capa lógica, ya que todas las actividades de la misma requieren en algún punto de su funcionamiento.

La integración del sistema ENKI-DB la completa la capa de presentación, que constituye la parte "visible" del sistema, ya que es la interfase con la cual el usuario final interactúa a través del sitio *web* implementado utilizando plantillas Smarty (Ohrt M., Zmievski A., 2005) (figura 4). Esta interfase permite al usuario acceder a los procesos de consulta, presentación de resultados, estadística y a la posibilidad de registrar secuencias obtenidas a partir de organismos de la biodiversidad colombiana dentro del sistema.

Disponibilidad del sistema. ENKI-DB es accesible en Internet en la siguiente dirección: http://bioinf.ibun.unal.edu.co/enkidb/ .

RESULTADOS Y CONCLUSIONES

La implementación de ENKI-DB como plataforma de información prototipo que cruza las bases de datos SIB y SPICA® con las moleculares EMBL y UNIPROT, da como resultado un sistema que permite la búsqueda de secuencias de ácidos nucleicos y proteínas relacionadas con especies de organismos propios de la biodiversidad de colombiana, a partir del nombre de la especie o de una secuencia problema. El cruce de estas bases de datos provee a la comunidad científica nacional e internacional de un sistema que permite relacionar hasta el momento, 10 808 especies propias de la biodiversidad colombiana con 96 337 registros de la base de datos UNIPROT, que comprende dos fuentes de información: UNIPROT Swissprot y UNIPROT Trembl (tabla 1), vinculados a 3 062 especies (28.33% del total, en promedio 31,46 entradas por especie) es decir, que 7 746 de dichas especies (71.67%) no cuentan, hasta la fecha, con información molecular registrada en UNIPROT.

También se pueden relacionar con estas 10 808 especies pertenecientes a la diversidad colombiana registradas hasta ahora, 1 880 414 registros en la base de datos EMBL, la cual comprende 9 fuentes de información: emblrelest, emblrelgss, emblrelhtg, emblrelmain, emblnew, emblcontigs, emblwgsrelease, emblwgsnew y embltpa, vinculados a 2 930 especies (28.11%), es decir que 7 878 de dichas especies (72.89%) no cuentan, hasta la fecha, con información molecular registrada en EMBL.

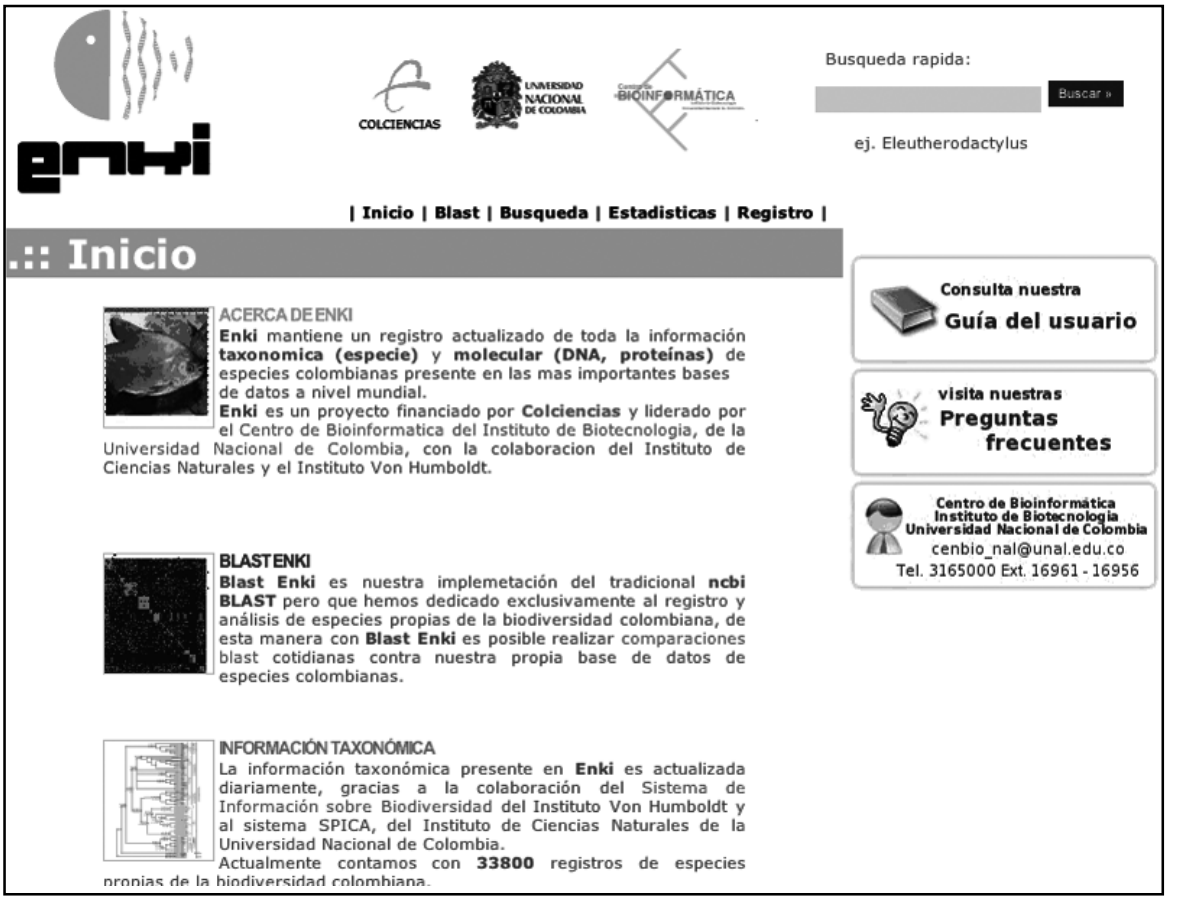

Figura 4. Interfase de usuario del sistema ENKI-DB

Es importante anotar que los registros encontrados en las bases de datos moleculares corresponden a organismos de la misma especie a los reportados dentro de las base de datos de biodiversidad colombiana, pero no fue posible obtener de las bases de datos moleculares el origen del espécimen, lo cual es una limitación para cierto tipo de estudios biológicos como los epidemiológicos, biogeográficos o los de filogenia, aunque los datos del sistema sirvan con referencia. Este hecho es un problema que se presenta a nivel mundial ya que, al revisar la mayoría de estas bases de datos moleculares, no se encuentra el sitio de origen de las muestras de las que fueron obtenidas las secuencias, probablemente debido a inconvenientes en las regulaciones de cada país o a la omisión por parte de los diseñadores de las bases de datos.

La implementación de ENKI-DB, facilita a los investigadores el registro del origen de las secuencias, un sistema de registro de muestras, el cual fue probado con los datos de secuencia de los genes 16RNA ribosomal y citocromo 2B de algunas especies del género *Eleutherodactylus*, los cuales fueron almacenados a manera de prueba dentro del prototipo de ENKI-DB. Este prototipo fue creado con el objeto de animar a todas las personas involucradas en el conocimiento de la biodiversidad a explicitar el origen exacto de los materiales con los que adelantan sus investigaciones a nivel molecular y promover la discusión sobre el tema.

ENKI-DB como prototipo puede ser adaptado fácilmente para recibir información de otras bases de datos tanto del tipo de SPICA® y SIB como de tipo molecular, ya que con pequeñas modificaciones y un mínimo de acuerdo con los responsables de cada base de datos, es posible incorporar la nueva información dentro del sistema y cruzar los datos automáticamente con los datos moleculares.

Tabla 1. Número de registros moleculares enlazados por el sistema ENKI- DB

Base de datos	Sección	Número de registros	Número de registros totales	Número de especies con registros en ENK-DB I	Registros asociados a especies en ENKI- DB	Representación mundial
UNIPROT	Uniprot Swissprot	195058	2'701.944	3062	96337	0,11%
	Uniprot Trembl	2506886				
EMBL	emblrelest	31990232				
	emblrelgss	13405392				
	emblrelhtg	78979				
	emblrelmain	6867693				
	emblnew	4088271				
	emblcontigs	493341				
	emblwgsrelease	9107071				
	emblwgsnew	3057592				
	embltpa	4647				

ENKI-DB cuenta con información completamente actualizada de todas las especies registradas en las bases de datos taxonómicos SPICA® y SIB, conjuntamente con sus datos moleculares (ADN y proteínas) proporcionando, de manera gratuita, información a investigadores e interesados en el área, quienes pueden consultar de manera actualizada la información molecular y taxonómica de las especies colombianas de su interés de una manera efectiva, llenando el vacío creado por los desarrollos independientes en cada área del conocimiento.

A pesar de que ENKI-DB ha sido implementado como un medio que contribuye al conocimiento de la biodiversidad colombiana, no significa que funcione exclusivamente para Colombia o para una región determinada, sino que es un modelo que puede ser fácilmente implementado por cualquier país o incluso como un sistema de integración de datos taxonómicos y moleculares a nivel mundial.

Dirección futura. Se están desarrollando nuevas características al sistema que mejoran tanto su experiencia de uso, como el tipo de análisis e información que provee. En este sentido se encuentra en fase de estudio la posible implementación en ENKI-DB de un sistema de análisis similar a ENSEMBL (Hubbard, et ál., 2005), a partir del cual se pueda generar información novedosa en genómica comparativa y funcional con referencia a la biodiversidad colombiana. Adicionalmente se está en el proceso de integración del paquete de análisis EMBOSS al módulo de visualización ya existente, de tal manera que sea posible realizar, por ejemplo, una búsqueda de microsatélites o una identificación de marcos de lectura abiertos directamente desde la página de resultados.

Otro aspecto fundamental que se está abordando es que el flujo de información se pueda dar en las dos vías, tanto desde ENKI-DB hacia SIB/SPICA®, como desde estos dos últimos hacia ENKI-DB.

AGRADECIMIENTOS

Este proyecto no hubiese sido posible sin el apoyo financiero de COLCIENCIAS y de la División de Investigación de Bogotá de la Universidad Nacional de Colombia y la colaboración e incondicional

ayuda del Instituto von Humboldt y del Instituto de Ciencias Naturales de la Universidad Nacional de Colombia, a cuyos miembros expresamos nuestros más sinceros agradecimientos.

BIBLIOGRAFÍA

Apweiler, R.; Bairoch, A.; Wu, Ch.; Barrer, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; López, R.; Magrane, M.; Martin, MJ.; Natale, DA.; O'Donovan, C.; Redaschi, N.; Yeh, L.S. 2004. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res*. 32: D115-D119.

Berman, H.M.; Henrick, K.; Nakamura, H. 2003. Announcing the worldwide Protein Data Bank. *Nature Structural Biology.* 10 (12), 980.

Centro de Investigación y Gestión de la Biodiversidad en Costa Rica. 2006. [en línea], disponible en: www.inbio.ac.cr. Fecha de consulta: 10/09/2006. Fecha de actualización: 11/08/2006.

Cochrane, G.; Aldebert, P.; Althorpe, N.; Andersson, M.; Baker, W.; Baldwin, A.; Bates, K.; Bhattacharyya, S.; Browne, P.; Van Den Broek, A.; Castro, M.; Duggan, K.; Eberhardt, R.; Faruque, N.; Gamble, J.; Kanz, C.; Kulikova, T.; Lee, C.; Leinonen, R.; Lin Q.; Lombard, V.; López, R.; McHale, M.; McWilliam, H.; Mukherjee, G.; Nardote, F.; Pastor, M.P.; Sobhany, S.; Store, P.; Tzouvara, K.; Vaughan, R.; Wu, D.; Zhu, W.; Apweiler, R. 2006. EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Research.* 34, D10-D5.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad en México (CONABIO). 2006. [en línea], disponible en: http://www.conabio.gob.mx/ . Fecha de consulta: 10/09/2006. Fecha de actualización: 27/04/2006.

European Bioinformatics Institute (EMBL) 2006. [en línea], disponible en: http://www.ebi.ac.uk/embl/. Fecha de consulta: 10/09/2006. Fecha de actualización: 09/09/2006.

Expasy Proteomics Server (TrEMBL). 2006. [en línea], disponible en: http://ca.expasy.org/. Fecha de consulta: 09/11/2006. Fecha de actualización: 10/08/2006.

Galperin, M.Y. 2006. The Molecular Biology Database Collection: 2006 update. *Nucleic. Acids Res*. 34: D3-D5.

Georgetown University Medical Center, Protein Information Resource (PIR). 2006. [en línea], disponible en: http://pir.georgetown.edu/. Fecha de consulta: 10/07/2006. Fecha de actualización: 05/07/2006.

Global Biodiversity Information Facility. 2006. [en línea], disponible en: http://www.gbif.org/. Fecha de consulta: 10/09/2006.

Hubbard, T.; Andrews, D.; Caccamo, M.; Cameron, G.; Chen, Y.; Clamp, M.; Clarke, L.; Coates, G.; Cox, T.; Cunningham, F.; Curwen, V.; Cutis, T.; Down, T.; Durban, R.; Fernández-Suárez, X.M.; Gilbert, J.; Hammond, M.; Herrero, J.; Hotz, H.; Howe, K.; Iyer, V.; Jekosch, K.; Cari, A.; Kasprzyk, A.; Keefe, D.; Keenan, S.; Kokocinsci, F.; London, D.; Longden, I.; McVicker, G.; Melsopp, C.; Meidl, P.; Potter, S.; Proctor, G.; Rae, M.; Rios, D.; Schuster, M.; Searle, S.; Severin, J.; Slater, G.; Smedley, D.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Storey, R.; Trevanion, S.; Ureta-Vidal A.; Vogel, J.; White, S.; Woodwark, C.; Birney, E. 2005. Ensembl 2005. *Nucleic Acids Res*. 33, D447-D453.

Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*. 34, D354- 357.

McGinnis, S.; Madden, T.L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 32, W20-W25.

National Center for Biotechnology Information. GENBANK. 2006. [en línea], disponible en: http://www.ncbi.nlm.nih.gov/Genbank/index.html. Fecha de consulta: 26/09/2006. Fecha de actualización 04/10/2006.

National Institute of Genetics of Japan. DDBJ. 2006. [en línea], disponible en: http://www.ddbj.nig.ac.jp/ . Fecha de consulta: 29/09/2006. Fecha de actualización 26/09/2006.

Ohrt, M.; Zmievski, A. 2005. *Smarty manual*. New Digital Group, 179 p.

Rice, P.; Longden, I.; Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics.* 16(6), 276-277.

Sistema de Información Biótico Ambiental. (SPICA). 2006. [en línea], disponible en: http://www.spica.unal.edu.co/. Fecha de consulta: 10/09/2006.

Sistema de Información Integrado Taxonómico de Norteamérica. 2006. [en línea], disponible en: www.itis.usda.gov. Fecha de consulta: 21/11/2006. Fecha de actualización: 19/11/2006.

Sistema de Información sobre Biodiversidad, (SIB) 2006. [en línea], disponible en: http://

www.siac.net.co/. Fecha de consulta: 29/11/2006. Fecha de actualización: 26/11/2006.

Sistema de Recuperación de Secuencias, (SRS) 2006. [en línea], disponible en: http://srs.ibun.unal.edu.co:8080/srs81/. Fecha de consulta: 10/09/2006. Fecha de actualización: 09/09/2006.

Species 2000. 2006. [en línea], disponible en: http://www.sp2000.org/. Fecha de consulta: 25/09/2006. Fecha de actualización: 20/09/2006.

The Biological Collection Access Service. 2006. [en línea], disponible en: http://www.biocase.org/. Fecha de consulta: 10/09/2006. Fecha de actualización: 03/07/2006.

Tree of Life web Project. 2005. [en línea], disponible en: http://tolweb.org. Fecha de consulta: 10/09/2006.

UNIPROT. [en línea], disponible en: http://www.pir.uniprot.org/index.shtml. Fecha de consulta: 10/05/2006.