Artículo de revisión

Aplicación de la inteligencia artificial en la bioinformática, avances, definiciones y herramientas*

Artificial intelligence application on bioinformatics, progress, definitions, and tools

Simon Orozco Arias, Jeferson Arango López*****

**Estudiante Ingeniería de Sistemas y Computación, Universidad de Caldas

***Ingeniero de Sistemas y Computación, Mg. Ingeniería Computacional, Universidad de Caldas

Resumen

La bioinformática es la una disciplina relativamente nueva la cual ayuda con el descubrimiento de información biológica a través de la implementación de técnicas computacionales (López-Gartner et al. 2015). Esta unión surge debido a la problemática dada en fenómenos tan complejos como la genética, la simulación del efecto de medicinas, la predicción de enfermedades, etc. Todas estas situaciones manejan gran cantidad de información y variables, de allí surge la necesidad de apoyarse con las nuevas tecnologías. Pero hay circunstancias en las que ni las mejores plataformas tecnológicas pueden encontrar respuestas en un tiempo prudente, es aquí donde se hace fundamental el uso de herramientas, técnicas, frameworks y metodologías propias de la inteligencia artificial para optimizar la mayor cantidad de procesos, reduciendo el tiempo y el gasto computacional que provocan el manejo de esta información. En el siguiente artículo se desarrolla un estado del arte de las mejores formas de la aplicación de la inteligencia artificial en la bioinformática encontrada en la literatura.

Palabras clave: Bioinformatica, datos biológicos, inteligencia artificial, redes neuronales.

Abstract

Bioinformatics is a relatively new discipline which contributes discovery of biological information through implementation of computer techniques (Lopez-Gartner et al. 2015). This union surges due to problems found in complex phenomena, such as genetics, medicine effect simulation, disease prediction, etc. All of these situations involve a great amount of information and variables, leading to the need of support using new technologies. Notwithstanding, there are circumstances where even the best technologies are unable to find answers within a prudent time, it is here where it becomes necessary to use tools, techniques, frameworks and methodologies involved in artificial intelligence, in order to optimize the greater number of processes, by reducing time and computer expenses caused by managing such information. The following article develops a state-of-the-art of the best methods for application of artificial intelligence in bioinformatics found in literature.

Keywords: Bioinformatics, artificial intelligence, Artificial Neuronal Networks, Biological Data.

*Investigación adscrita al grupo de investigación GITIR, Universidad de Caldas

Recibido: 08/01/2016
Revisado: 13/03/2016
Aceptado: 01/12/2016

Correspondencia de autor:

simon.orozco.arias@gmail.com

jeferson.arango@ucaldas.edu.co

Cómo citar:

Orozco, S., Arango, J. (2016) Aplicación de la inteligencia artificial en la bioinformática, avances, deficiniones y herramientas. *UGCiencia* 22, 159-171.

Introducción

Gracias al crecimiento exponencial de las tecnologías de la información como los clúster[1], las grid[2] y la nube[3] y de los modelos aplicables a ellas como la inteligencia artificial y la minería de datos, además de la paralización de procesos y la accesibilidad a la información científica mundial, se están creando cada vez más, nuevos y mejores análisis de la información y técnicas adaptativas con la habilidad de aprender, con el fin de dejar a un lado la sociedad de la información para adentrarse en la sociedad del conocimiento.

Bioinformática: una ciencia con auge creciente

La bioinformática en su definición se caracteriza como el estudio de la información biológica a partir de la teoría de la información, la computación y las matemáticas, (Lahoz-Beltrá, 2010). Además es una nueva disciplina dentro de la biología, donde las herramientas de la computación tienen una función primordial y si bien algunos restringen el rango de estudio de la bioinformática al manejo y análisis de bases de datos biológicas principalmente de secuencias, también podría atribuírsele un sentido más amplio, como la fusión de las técnicas computacionales con el entendimiento y apreciación de datos biológicos, el almacenamiento, recuperación, manipulación y correlación de datos procedentes de distintas fuentes.

Una forma de graficar las secuencias obtenidas por el manejo de estos grandes volúmenes de datos es la bioinformática estructural. Según (Martí & Turjanski, 2009) consiste en realizar una simulación de comportamiento biomolecular, principalmente proteínas y sus entornos. La simulación puede establecer la capacidad de que una droga si puede tratar ciertas enfermedades, permitiendo así con más rapidez el descubrimiento y proceso óptimo de estas medicinas, por lo tanto es entonces la disciplina de la ciencia que se dedica al análisis del ADN y ARN para ordenar la información generada mediante experimentos y la aplicación de estos métodos para resolver problemas de índole biológico y así generar nuevo conocimiento. La bioinformática estructural ha tenido grandes aportes significativos, por ejemplo usando aminoácidos y ácidos nucleicos se han desarrollado modelos estructurales a partir de información obtenida por técnicas de rayos X y RMN[4](Resonancia Magnética Nuclear). Gracias a los avances en la secuenciación, como en la cristalografía a gran escala se crean nuevas oportunidades de usar proteínas en la búsqueda de fármacos.

Según (Castellanos, Ortiz, Nápoles, & Cáceres) se ha caracterizado en la bioinformática varios enfoques importantes, entre los cuales se encuentra el campo del código genético standard (CGS), el cual no es el resultado de un proceso de asignación aleatorio de aminoácidos a codones, sino todo lo contrario, se determinan regularidades estructurales que lo distinguen y que tiene profundas consecuencias en las propiedades de las secuencias biológicas actuales y en sus patrones evolutivos, revelan también el resultado de la acción de ciertas leyes.

La bioinformática ha tenido una tarea importante en la validación o invalidación de estas teorías que intentan explicar regularidades visibles en el CG (Código Genético) así como en la evaluación del efecto de estas características estructurales en las secuencias biológicas actuales por medio de sesgo en el uso de codones, uso de aminoácidos y tasa de sustituciones sinónimas

1. Cluster: Conjunto de equipos de cómputo (no necesariamente con hardware y software homogéneos) unidos a través de una red de datos de alta velocidad. (Meza Martínez & Uribe Hurtado, 2013)
2. Grid: Infraestructura tecnológica en la cual se conectan y se comunican múltiples equipos de cómputo generalmente separados geográficamente, con el fin de compartir recursos. (Dong & Akl, 2006)
3. Nube: es un modelo diseñado para permitir acceso ubicuo a la red bajo demanda a un conjunto de recursos informáticos compartidos configurables. (Mell & Grance, 2011)

4. Resonancia Magnética Nuclear: separación de los estados de los espines nucleares en presencia de un campo magnético intenso. (Blancas et al., 2010)g

o no sinónimas. Otros aportes importantes de la bioinformática, según Castellanos et al, 2005, son:

- Estimación de cuán óptimo es el código genético como filtro de errores de acuerdo con determinadas propiedades mediante simulación numérica o por vía analítica.
- Reconstrucciones de escenarios primitivos cuando surge y se desarrolla el código genético.
- Simulación numérica de la interacción evolutiva entre mensaje y código.
- Evaluación de cuán distante son los códigos más óptimos con relación al código genético standard, en estos trabajos se emplean algoritmos de optimización como simulated annealing[5] o algoritmos genéticos entre otros.
- Simulaciones de las fuerzas selectivas o neutrales que están detrás de los cambios de reglas de asignación que han dado lugar a las variantes conocidas actualmente del código genético.

Así como se aplicó en CGS la bioinformática también ha tenido grandes roles como lo definido en (Meléndez-Herrada, Ramírez, Sánchez Dorantes, & Cervantes, 2010) donde se habla del caso de la epidemia mundial del virus de la influenza A (H1N1) en el cual se estudiaron agentes infecciosos desde el punto de vista genético, evolución, propiedades anti higiénicas, etc. Analizando la información se creó una respuesta inmune donde se aplicó el programa inmune Epitope Database[6]. En este programa se clasifican los epítopos de tipo B (respuesta anticuerpos) y de tipo T (respuesta celular), como resultado importante de este análisis bioinformático fue que el 35% de los epítopos del nuevo virus eran reconocidos por la respuesta de los anticuerpos mientras que el 67% son reconocidas por células, donde se notó que el resultado es muy variable para las proteínas HA y NA. Gracias a este tipo de análisis se logró el desarrollo de la vacuna de la influenza universal. Comparado con anteriores pandemias de influenza ahora se tiene un tiempo de respuesta más rápido en conocer nuevas características y antigénicas de estos virus, basados en la genómica y en el análisis bioinformático.

Cuando se habla de bioinformática se piensa en sus aplicaciones y las ventajas que ha tenido su utilización, pero para lograr estos objetivos también se debe pensar en plataformas para su correcto funcionamiento, como la propuesta por Castillo et al., 2015, en donde existe una aproximación a un Web Service con una arquitectura basada en la plataforma (GITIRBio) que actúa como un sistema front-end distribuido para procesamiento autónomo y asistido por tuberías paralelas bioinformáticas, donde se utiliza para la validación del uso de múltiples secuencias, Así esta plataforma permite escalabilidad o mejor aún repositorios semánticos de genes para las anotaciones de búsqueda.

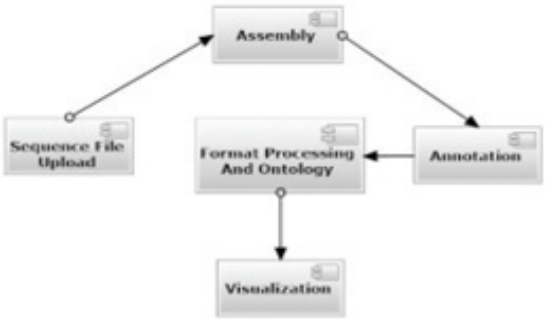
Esta idea surge para cubrir la necesidad de gran cantidad de científicos que no se han familiarizado con la evolución de la informática de alto rendimiento en línea de comandos sobre la paralelización, donde muchas de las entidades no proporcionan un sistema integrado, guiado y ayudado a la interacción. Este sistema se compone de 2 partes, el módulo de comunicación y el de procesamiento, donde el módulo de comunicación adquiere la petición de usuario a través de la interfaz de usuario y luego lo entrega al módulo de proceso el cual es un envoltorio que ejerce en todas las interacciones con las herramientas bioinformáticas como gestión, configuración de parámetros, la gestión de la

5. simulated annealing: fue propuesto como un algoritmo que está basado sobre la analogía entre el tratamiento de sólidos y el problema de resolver problemas de optimización combinacional (Pham & Karaboga, 2012)
6. Inmune Epitope Database programa que proporciona un catálogo experimental de células epítopes B y T caracterizadas (Iqdour & Zeroual, 2006)

producción, el manejo de errores y notificación de envió.

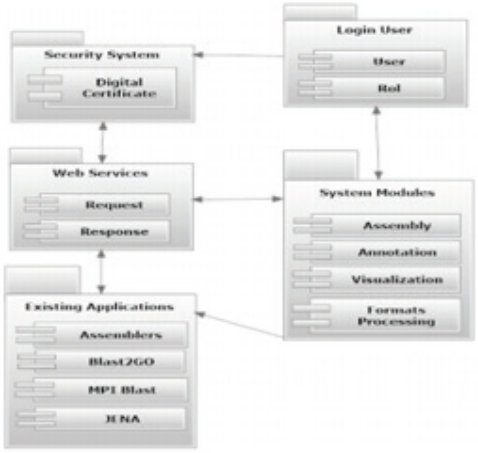
En la ilustración 1 muestra la relación lógica de los módulos principales del sistema utilizado en el procesamiento y en el caso de uso general.

Ilustración 1 . Relación de los módulos principales en GitirBio.



El objetivo de esta arquitectura es generar un registro de seguridad para todos los datos procesados por las fases en las tuberías, donde estos datos se refieren a la secuencia de montaje que puede ser realizado por un número específico de funciones, los cuales dependen de la anotación y la visualización de la secuencia de ensamblado de acuerdo con el método seleccionado. El comportamiento del procesamiento se muestra en la Ilustración 2.

Ilustración 2.Comportamiento del procesamiento en GitirBio



Fue

Propósitos

Según Juan M. Corchado, 2015, proponen que en el área de la inteligencia artificial distribuida para el descubrimiento de conocimiento en bioinformática, existen muchos propósitos, pero en la mayoría se analizan diferentes aspectos biológicos y se simulan dichos procesos o comportamientos en un sistema natural saludable. Algunos de los propósitos son: 1) "Bladder Carcinoma Data with Clinical Risk Factors and Molecular Markers: A Cluster Analysis" propone la hipótesis de que el uso de datos clínicos e histopatológicos es muy útil para manejar tratamientos de cáncer de vejiga invasivo no muscular (NMIBC). Los autores usan minería de datos[7] en un clúster médico con el fin de analizar pacientes y dividirlos en varios grupos usando nuevas técnicas de diagnóstico de NMIBC. Los pacientes fueron categorizados acorde a las características médicas y a su comportamiento biológico. (Redondo-Gonzalez et al., 2015). 2) Los autores de "A Linear-RBF Multikernel SVM to Classify Big Text Corpora" usan técnicas de minería de datos basadas en clasificadores. Para reducir el costo computacional usan computadores multikernel que soportan vectores (SVM) con parametrización automática, con el fin de paralizar procesos y reducir la cantidad de datos a procesar. (Romero, Iglesias, & Borrajo, 2015). 3) El paper "Gene Knockout Identification Using an Extension of Bees Hill Flux Balance Analysis" propone un modelo en el cual usan una extensión del análisis del flujo balanceado de las colmenas de abejas (BHFBA) integrado con el framework OptKnock[8] y Hill Climbing algorithm[9] para extraer mayor cantidad de información de un gen especificado en un metabolismo determinado. (Choon et al., 2015). 4) En "Using the eServices Platform for

7. Minería de datos proceso de descubrimiento de nuevas y significaciones relaciones patrones y tendencias al examinar grande cantidades de datos (Riquelme, Ruiz, & Gilbert, 2006)
8. framework OptKnock Usado para predecir estrategias knockout de genes destinados a la sobreproducción de un metabolito deseado. (Choon et al., 2014)
9. Hill Clinbing es un algoritmo iterativo, que en cada iteración usa la solución actual para determinar la candidatura de una nueva solución. (Burke & Bykov, 2008)

Detecting Behavior Patterns Deviation in the Elderly Assisted Living: A Case Study," dado por Marcelino et al., 2015, los autores usan una plataforma de eService para detectar cualquier variación en el comportamiento estudiado y puede predecir situaciones peligrosas. El sistema fue modelado bajo la metodología CRISP-DM, además de usar técnicas exhaustivas de búsquedas de minería de datos como árboles de decisión, clústeres, o curvas en orden para validar los resultados.

Otro de los propósitos encontrados en la actualidad es la de simplificar la complejidad y cantidad de datos analizados obtenidos a partir de la bioinformática, usando técnicas de la Inteligencia Artificial con el fin de mejorar y optimizar los procesos de decisión.

El anterior es el caso descrito en Guillermo Roberto Salarte Martínez, 2012, en el diagnóstico clínico de enfermedades cardiovasculares.

Los autores plantean la problemática de la complejidad de clasificar los datos a través del método de redes bayesianas, ya que cada nodo del grafo representa una variable que compone el dominio, aunque esta no tenga relación directa con la tarea especificada. Por lo tanto se propone unir las ventajas de esta técnica con las ventajas de los árboles de decisión.

De esta forma se obtiene un modelo hibrido que está formado por dos fases:

La primera etapa consiste en la preselección de nodos y construcción de la red, es decir que, a partir de los datos que se encuentran en una base de datos, esta se encarga de la selección y clasificación de un subconjunto de nodos para mejorar la capacidad predictiva de la red

La segunda fase consiste en la construcción de la red bayesiana[10] a partir del subconjunto de variables seleccionadas en la etapa previa, aplicando el método de aprendizaje de redes bayesianas (algoritmo de probabilidades).

Herramientas

Frameworks

En lo expuesto por Miguel P. Rocha, 2009, se discute una problema que surge debido a la expansión de la bioinformática, y es que al manejar cantidades excesivamente grandes de datos se requiere indudablemente de capacidad de computo, de algoritmos altamente complejos y paralelos, de procesamiento en CPUS, GPUS u otros tipos de procesadores y no siempre estas tecnologías son lo suficientemente homogéneas o compatibles para el uso en la bioinformática. Por otro lado las personas que realizan los estudios de esta índole son en su mayoría biólogos o científicos los cuales no tienen cualidades para programar, algo que es de fundamental importancia para analizar los datos recolectados en los estudios y dar respuestas acordes a estos.

Los autores proponen una herramienta basada en una nueva, abierta, libre y cumpliendo con una arquitectura documentada llamada Biomniverso, la cual es fácil de usar, tiene una GUI amigable y es fácilmente configurable, sin escribir código.

Esta herramienta se compone de dos partes fundamentales, el kernel, llamado Omega, que suple los servicios específicamente adaptados para permitir la adición de nuevas funcionalidades bioinformáticas por medio de plugins. Por otro lado está la interfaz, llamada Brigid, la cual permite a los científicos de laboratorio trabajar con diferentes procesos con un esfuerzo mínimo y ejecutar scripts mediante una GUI amigable y de fácil uso.

Otro framework es propuesto por HAMDI-CHERIF, 2010, allí se propone un marco de trabajo de tres niveles para su uso en la bioinformática, estos niveles son caracterización

10. Red Bayesiana es una construcción matemática que representa una articulación de la distribución probabilística entre un conjunto de variables. (Tsamardinos, Brown, & Aliferis, 2006)

de programas de inteligencia libre, basados en inteligencia artificial y programas de control de inteligencia. Cada uno de estos niveles es un mapeo directo del desarrollo histórico correspondiente al entendimiento.

La intención del autor fue unificar una máquina que aprende con teorías de control en bioinformática. Se habla también sobre dos enfoques que ha tenido esta ciencia, en los cuales se usan programas estándares, heurísticos y libres como manejadores de bases de datos, seguido de programas basados en inteligencia artificial limitada. Otro enfoque propone que además de lo descrito anteriormente se adicione una acción de control inteligente.

Estos tres enfoques pueden ser vistos como niveles de comprensión con un grado creciente de complejidad. La inteligencia artificial también es usada para tratar temas bioinformáticos muy específicos, los cuales no tienen solución bajo otros enfoques, como ejemplo se expone lo propuesto por Pokkuluri Kiran Sree, 2014, donde se habla acerca de la solución que se tiene para muchas de las situaciones presentadas en el proceso investigativo usando autómatas celulares[11] (CA).

En el artículo anterior se describe un autómata celular como un conjunto de células con un número finito de estados y está definido de la siguiente forma (Definición tomada de Pokkuluri Kiran Sree 2014):

CA is defined a four tuple <G, Z, N, F>

Where G -> Grid (Set of cells)

Z -> Set of possible cell states

N -> Set which describe cells neighborhoods

F -> Transition Function (Rules of automata)

Cada célula dentro del grid del CA, tiene un elemento de memoria (D flip flop) con algún integrado lógico (XOR o XNOR), cada célula es actualizada en cada ciclo de reloj, las transiciones de ellas dependen de sus vecinos, tal y como lo define (Pokkuluri Kiran Sree 2014). La tabla 1 grafica las reglas de los vecinos y el siguiente estado relacionado a ellas.

Tabla 1. Reglas de vecinos.

Possible Combinations	111	110	101	100	011	010	001	000
Binary Equivalent of Rule-254 (Next State)	1	1	1	1	1	1	1	0

Fuente: (HAMDI-CHERIF, 2010)

Con este enfoque los autores demuestran que pueden reconocer las regiones promotoras en cadenas de proteínas, además de predecir la estructura de dicha proteína.

Uno de los problemas a los que se enfrenta la bioinformática es a la hora de la utilización de las múltiples herramientas desarrolladas, al ser tan diversas y con propósitos tan diferentes se hace indispensable la automatización de la ejecución de estas herramientas, sin perder demasiado tiempo configurándolas, según (Barraza, Salazar, Cuesta-Astroz, & Restrepo) se puede llegar a una aproximación usando los flujos de trabajo. Los autores proponen una aplicación escrita en perl[12] la cual ejecuta en forma sucesiva y controlada cada una de las herramientas implicadas en el análisis bioinformático.

Según lo expuesto por Pelta (2013) plantea que debido al surgimiento de problemas en bioinformática, se han creado algoritmos utilizando técnicas heurísticas para resolverlos

11. Autómatas Celulares sistemas espaciales dinámicos muy simples en los que el estado de cada celda depende de los estados previos de las celdas vecinas. (Aguilera Benavente, 2006)

12. Perl lenguaje de programación orientado a la extracción de información desde archivos de texto. (Hammond, 2008)

y ya que la complejidad computacional es demasiado alta se ha visto la necesidad del uso de lógica difusa para dar lugar a una herramienta robusta y flexible la cual es adaptada para el campo de la bioinformática, el método desarrollado se denomina Fuzzy Adaptive Neighborhodd Search (FANS) y es esencialmente una herramienta de optimización basada en búsquedas por entornos que incorpora como elementos novedosos, la utilización de una valoración difusa de las soluciones y la utilización de varios operadores en este proceso. La manipulación de FANS hace que su comportamiento sea similar a otros métodos de búsqueda por entornos, lo que permite plantear que fans es un "framework" simple, además de una herramienta capaz de obtener soluciones razonablemente buenos y con poco esfuerzo computacional, pero tiene la característica de detectar mejores soluciones para patrones grandes que para pequeños, por lo tanto este método matemático se utiliza para resolver satisfactoriamente cada instancia de prueba con esfuerzo reducido.

En la investigación de Hernández (2008) se proponen algunas estrategias nuevas para el análisis en la bioinformática como el desarrollo de herramientas que usan la proyección al diseño y la generación de sistemas eficientes de almacenamiento y nuevos modelos para la comparación y análisis de las distintas clases de datos biológicos, rápidos y confiables desde el punto de vista estadístico, como es el caso del algoritmo BLAST[13]. También la utilización de tecnologías de alto rendimiento en investigación biológica, lleva a que las actuales estrategias de análisis tengan un proceso de adaptación o la creación de nuevos desarrollos, para aprovechar de mejor manera los recursos disponibles, como ocurre en el caso de las nuevas metodologías de secuenciación. Ya basadas actualmente en la química de sangre y su asignación de bases, la cual se caracteriza por longitudes de lectura cortas y altos porcentajes de error en las secuencias, lo que requiere de nuevas formas de asignación de las bases, de ensamblaje de las secuencias y alteración de los métodos estadísticos para la determinación de puntajes de calidad. Gracias a estas estrategias los investigadores ya cuentan con un muestreo de más de 100 condiciones de ambientes diferentes utilizando aproximaciones metagenómicas, donde estos diseños implican cambios en el diseño de algoritmos para aprovechar el alto rendimiento de las máquinas para el manejo de grandes datos.

Enfoques o técnicas

Algunos problemas típicos en la bioinformática son la clasificación, la detección de patrones y la predicción, debido a que muchos de los procesos en esta área usan gran cantidad de datos, de los cuales se desea sacar conocimiento y en donde no todos los datos son relevantes para el caso estudiado, los tiempos de análisis y procesamiento pueden ser muy extensos y además requerir gran capacidad de computo.

Esta problemática se afronta directamente en Bonet, Rodríguez, García, & Grau (2012) donde se propone una nueva técnica basada en aprendizaje automático, la cual selecciona un clasificador según la situación, tal y como lo hacen las personas en el ambiente natural.

Los autores también hablan sobre algunos clasificadores existentes en la literatura como Bagging[14], Boosting[15] y Stacking; donde en esencia tienen dos partes importantes: selección de los clasificadores de base y elección de la forma de combinar las salidas.

Se escogieron 3 clasificadores de base: J48, red bayesiana y SVM. Como metaclasificador

13. Blast programa de búsqueda de similitud de secuencias que se pueden utilizar a través de una interfaz web o como una herramienta independiente (Johnson et al., 2008)

14. Bagging es una metodología para generar múltiples versiones de un sistema capaz de predecir. (Shinzawa, Jiang, Ritthiruangdej, & Ozaki, 2006)

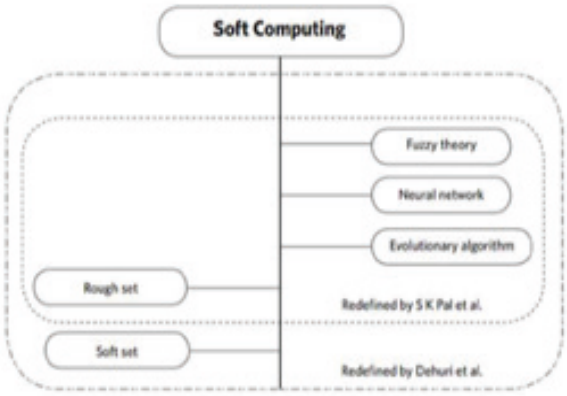
15. Boosting es un algoritmo de aprendizaje basado en conjuntos, el cual tiene el propósito de mejorar la precisión de la clasificación de forma iterativa. (Shinzawa et al., 2006)

se utilizó un MLP. Los autores hicieron un estudio del umbral para formar los grupos por clasificadores, en correspondencia con cada una de las bases. Para las diferentes bases utilizadas, los mejores resultados se alcanzaron con umbrales distintos como era de esperar, ya que las bases tienen características diversas. De manera general el valor del umbral osciló entre 0,6 y 0,9. Tal y como se nombra por Bonet et al., 2012g.

Para validar los resultados del modelo de combinación de clasificadores que se propuso, se comparó con Bagging, Boosting y Stacking[16], por ser los multiclasificadores más comúnmente utilizados en la literatura. En los casos de Bagging y Boosting se probaron tres clasificadores bases: J48[17], SVM[18] y MLP[19]. En el caso de Stacking se utilizó la misma topología usada para el modelo anteriormente propuesto.

En la literatura existen innumerables paradigmas para tratar información con origen biológico, además de maneras y herramientas para solucionar los problemas que ello conlleva, tal es el caso presentado por Hiwarkar & Iyer, 2013. Los autores proponen el uso de Soft Computing que se define como técnicas empleadas para solucionar problemas que manejan información incompleta, con incertidumbre e inexacta. A continuación se muestra una gráfica tomada del artículo anteriormente descrito donde se ve claramente los componentes de la Soft Computing.

Ilustración 4 Tomado de: (Hiwarkar & Iyer, 2013)



También hace énfasis en que la mayoría de las investigaciones se centran alrededor de procesos como el reconocimiento de patrones y la minería de datos para realizar tareas tales como la clusterización, clasificación, selección de características y la generación de reglas.

Uso de Redes Neuronales artificial en la bioinformática

Una red Neuronal Artificial (ANN) es un modelo informático capaz de capturar y representar relaciones complejas de entradas y salidas de manera similar al cerebro humano (Qian & Sejnowski, 1988; Tablada & Torres, 2009). Una ANN es capaz de aprender desde ejemplos y generalizar para encontrar una solución viable a una situación dada.

Lógica difusa en Bioinformática

Puede ser usaba muy fácilmente para implementar sistemas desde simples y pequeños hasta grandes y robustos. La lógica difusa[20] reduce la cantidad de pasos y simplifica la complejidad inherente al problema. El primer paso consiste en entender y caracterizar

16. Stacking es un método de clasificación que se caracteriza por el empleo de diferentes modelos, que combina las salidas usando un metaclasificador. (Kurczynski & Gawiser, 2010)
17. J48 es un árbol de decisión C4.5 para la clasificación que crea un árbol binario (Patil & Sherekar, 2013)
18. SVM método de aprendizaje de máquina supervisado que se utiliza ampliamente para problemas de clasificación y regresión. (Shamim, Anwaruddin, & Nagarajaram, 2007)
19. MLP es el modelo de red neuronal más común y es conocido como una red supervisada porque requiere la salida deseada ordenada para aprender (Iqdour & Zeroual, 2006)

20. Lógica difusa conocida por contemplar no sólo las opciones de verdadero y falso, sino también las múltiples variables de respuesta (Cañellas & Brage, 2006)

el comportamiento del sistema usando el conocimiento y la experiencia. También se puede usar para optimizar la minería de datos con el fin específico de mejorar el proceso de agrupamiento (Porras, Laverde, & Diaz, 2008).

Étnicas de algoritmos genéticos[21] en Bioinformática

Se trata de búsquedas aleatorias guiadas bajo los principios de la evolución de las especies. Provee soluciones para problemas multi objetivos, optimizando los requerimientos computacionales y brindando robustez (Hiwarkar & Iyer, 2013).

Rough Sets en la Bioinformática

Es una metodología muy nueva en el ámbito médico, es usada para descubrir dependencias entre datos, evaluar importancia de atributos, descubrir patrones en los datos, reducir redundancia en la información y atributos, reconocer y clasificar objetos, entre otros. Singularmente es usado para obtener reglas de las bases de datos; una de las principales ventajas es que crea reglas del tipo if-then (Hassanien, Milanova, Smolinski, & Abraham, 2008).

Particle Swarm Optimization (PSO)

Es un tipo de inteligencia colectiva en sistemas de agentes descentralizados. Están basados en las colonias de la naturaleza, tales como colonias de hormigas, bandadas de pájaros, colmenas de abejas, bacterias y microorganismos y se usa para optimizar procesos. Fue concebida imitando el comportamiento social de los humanos y al ser un sistema de agentes descentralizados no necesita grandes volúmenes de información para optimizar los procesos estudiados, por lo tanto solo se usa operaciones matemáticas simples entre los miembros del enjambre (Hassanien et al., 2008).

21. Algoritmos genéticos son utilizados para encontrar la combinación óptima de variables explicativas para un modelo multivariado tradicional (Parisi, Parisi, & Díaz, 2006)

Buenas prácticas en la Bioinformática

Como toda ciencia existen unas técnicas que garantizan los buenos procesos llevados en ellos y la bioinformática no es la excepción. Según Kelley & Rouchka (2007) se habla sobre unas de las muchas técnicas desarrolladas para el campo bioinformático. En dicho estudio se usaron técnicas para la investigación de la diabetes donde se determinó que es necesario primero para desarrollar categorías, herramientas y técnicas, como las nombradas a continuación:

- Proyectos de alineación de secuencia y técnicas, los cuales son muy nombrados en la literatura como una herramienta primaria para la investigación, la cual se podría incluir por parejas y secuencias múltiples como los son las búsquedas Blast. Esta técnica es utilizada para comparar ya sea ADN o las secuencias de aminoácidos de los organismos para determinar homología y generar relaciones filogenéticas entre ellos.

- Proyectos de expresión génica y técnicas donde se cita métodos para medir expresiones de genes en diferentes organismos y condiciones, gracias al análisis de microarrrays la cual se menciona con frecuencia en muchos estudios.

- Las bases de datos y técnicas de bases de datos donde su propósito puede ser ayudar en la investigación o ayudar a otros investigadores con sus trabajos.

Computación de altas prestaciones

Debido a las características de la bioinformática anteriormente mencionadas en este artículo, se debe usar plataformas tecnológicas robustas, con capacidad de cómputo y arquitecturas especializadas. Es por esto que esta ciencia se sostiene sobre la supercomputación de altas prestaciones (HPC) (Apon et al., 2010).

En la actualidad existen máquinas de alto rendimiento, las cuales se pueden detallar con características específicas en el top500 según Kogge & Dysart, 2011g, las cuales cuentan con recursos como procesadores de varios núcleos, tarjetas gráficas y redes de alta velocidad. La arquitectura que predomina en el mercado es la de clúster.

Para aprovechar en mayor medida esta plataforma tecnológica, en la actualidad se aplica el modelo de paralelización de procesos, programación concurrente y distribuida y muchos otros paradigmas (Minetti, 2012).

En la actualidad empresas como NVidia, tiene proyectos para crear arquitecturas, tanto de software como de hardware para ejecutar con mayor rapidez y eficacia aplicaciones bioinformáticas como Blast (Schmidt, 2010).

Debido a la gran cantidad de información que se genera día a día, la cual se encuentra en la escala de los exabytes, se deben buscar mecanismos para procesar e identificar más rápidamente información relevante en computación paralela. En los últimos años la tecnología GPGPU[22] ha tomado un gran auge. Es en este aspecto donde Amir, 2013, presenta una investigación en la cual formula una nueva forma de aplicar big data usando las GPUs, los autores afirman que la computación paralela es la característica clave de la tecnología big data[23] en la abstracción de varios niveles.

Conclusión

En la actualidad existen diferentes necesidades en el ámbito científico alrededor del procesamiento de datos biológicos, y las ciencias de la computación desarrollan un papel fundamental en el avance de generación de herramientas para la consecución de resultados con mayor rapidez y la misma o mayor fiabilidad de la que se tiene desde hace unas décadas.

Debido al avance computacional en el campo del hardware, es necesario construir de igual forma plataformas que cumplan con características especiales que obtengan el mayor beneficio posible de estas máquinas y mediante técnicas de paralelización, inteligencia artificial, machine learning, entre otras.

Es de esta forma que las investigaciones en las ciencias de la vida que deban obtener datos a partir de procesos informáticos se adhieren y permiten compartir conocimiento para así, llegar a un fortalecimiento futuro de la bioinformática.

Referencias bibliográficas

Aguilera Benavente, F. (2006). Predicción del crecimiento urbano mediante sistemas de información geográfica y modelos basados en autómatas celulares. *Geofocus*, 6, 81-112.

Amir, A. (2013). *Implementation of Bio-Informatics Applications on Various GPU Platforms*. Delft: Delft University of Technology.

Apon, A., Ahalt, S., Dantuluri, V., Gurdgiev, C., Limayem, M., Ngo, L., & Stealey, M. (2010). High performance computing instrumentation and research productivity in US universities. *Journal of Information Technology Impact*, 10 (2), 87-98.

Barraza, F., Salazar, G., Cuesta-Astroz, Y., & Restrepo, O. E. (2006). *Implementación de una arquitectura web para la ejecución de flujos de trabajo en bioinformática*. Tomado de: http://bibliotecadigital.univalle.edu.co/bitstream/10893/1609/1/inycompe_v8_n2_a4.pdf

Blancas, R. B. P., Cárdenas, M. R. J., Cerezo, R. P., Lozano, R. R., Gómez, B. T., & Haddad, J. L. (2010). Enfermedad humana por modelantes. Análisis de sustancias con espectrometría de resonancia magnética. *Cirugía plástica*, 20(3), 120-123.

22. GPGPU Utilización de las capacidades de las GPUs para propósitos generales (Amir, 2013)
23. Big Data concepto que abarca el almacenamiento de grandes cantidades de datos y de sus aplicaciones en la industria. (Provost & Fawcett, 2013)

Bonet, I., Rodríguez, A., García, M. M., & Grau, R. (2012). Combinación de clasificadores para bioinformática. *Computación y Sistemas*, 16, 191-201.

Burke, E. K., & Bykov, Y. (2008). *A late acceptance strategy in hill-climbing for exam timetabling problems. Paper presented at the PATAT Conference*. Montreal: Canadá.

Cañellas, A. J. C., & Brage, L. B. (2006). Lógica difusa: una nueva epistemología para las Ciencias de la Educación. *Revista de educación* (340), 995-1008.

Castellanos, M. S., Ortiz, C. M. M., Nápoles, O. C., & Cáceres, J. L. H. (2005) El código genético desde la perspectiva de la bioinformática. *Centro de Cibernética Aplicada a la Medicina, Instituto Superior de Ciencias Médicas de la Habana*. Tomado de: http://www.cecam.sld.cu/#t1

Castillo, L. F., López-Gartner, G., Isaza, G. A., Sánchez, M., Arango, J., Agudelo-Valencia, D., & Castaño, S. (2015). GITIRBio: A Semantic and Distributed Service Oriented-Architecture for Bioinformatics Pipeline. *Journal of Integrative Bioinformatics*, 12(1), 255.

Choon, Y. W., Mohamad, M. S., Deris, S., Chong, C. K., Omatu, S., & Corchado, J. M. (2015). Gene Knockout Identifiation Using an Extension of Bees Hill Flux Balance Analysis. BioMed research international. DOI: 10.1155/2738.

Choon, Y. W., Mohamad, M. S., Deris, S., Illias, R. M., Chong, C. K., Chai, L. E., Corchado, J. M. (2014). Diffrential bees flx balance analysis with optknock for in silico microbial strains optimization. *PloS one*, 9(7), e102744.

Dong, F., & Akl, S. G. (2006). Scheduling algorithms for grid computing: State of the art and open problems: Technical report. Ontario: Queen´s University.

Salarte, G; Castro, Y (2012). Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial. Tecnura, 16(33) pp. 35- 52.

Hamdi-Cherif, A. (2010). *Machine Learning for Intelligent Bioinformatics – Part 2 Intelligent Control Integration*. Recent advances in artificial intelligence, knowledge engineering and data bases. Tomado de: http://www.wseas.us/e-library/conferences/2010/Cambridge/AIKED/AIKED-52.pdf

Hammond, M. (2008). Programming for linguists: Perl for language researchers: John Wiley & Sons. DOI: 10.1002/9780470752234

Hassanien, A.-E., Milanova, M. G., Smolinski, T. G., & Abraham, A. (2008). *Computational intelligence in solving bioinformatics problems: Reviews, perspectives, and challenges Computational Intelligence in Biomedicine and Bioinformatics*. New York: Springer.

Hernández, E. B. (2008). Bioinformática: una oportunidad y un desafío. *Revista Colombiana de Biotecnología*, 10(1), 132-138.

Hiwarkar, T. A., & Iyer, R. S. (2013). New Applications of Soft Computing, Artifiial Intelligence, Fuzzy Logic & Genetic Algorithm in Bioinformatics. *International Journal of Computer Science and Mobile Computing,* 2 (5) pp. 202-207.

Iqdour, R., & Zeroual, A. (2006). The Multi-Layered perceptrons neural networks for the prediction of daily solar radiation. *International Journal of Signal Processing*, 3(1), 24-29.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, 36(2), W5-W9.

Corchado, J; Bichindaritz, I y Paz, J (2015). Distributed Artificial Intelligence Models for Knowledge Discovery in Bioinformatics. *Biomedical Research International*. doi: 10.1155/2015/846785.

Kelley, R., & Rouchka, E. C. (2007). Bioinformatics Techniques Used in Diabetes Research. Kentucky: University of Louisville.

Kogge, P. M., & Dysart, T. J. (2011). Using the TOP500 to trace and project technology and architecture trends. Paper presented at the Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis

Kurczynski, P., & Gawiser, E. (2010). A Simultaneous Stacking and Deblending Algorithm for Astronomical Images. *The Astronomical Journal,* 139(4), 1592.

Lahoz-Beltrá, R. (2010). Bioinformática: Simulación, vida artificial e inteligencia artificial, Madrid: Ediciones Díaz de Santos.

López-Gartner, G., Agudelo-Valencia, D., Castaño, S., Isaza, G. A., Castillo, L. F., Sánchez, M., & Arango, J. (2015). *Identification of a Putative Ganoderic Acid Pathway Enzyme in a Ganoderma Australe Transcriptome by Means of a Hidden Markov Model*. In 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. Springer International Publishing.

Marcelino, I., Lopes, D., Reis, M., Silva, F., Laza, R., & Pereira, A. (2015). Using the eServices platform for detecting behavior patterns deviation in the elderly assisted living: case study. *BioMed research international*. DOI: 10.1155/2738.

Martí, M. A., & Turjanski, A. A. (2009). La bioinformática estructural o la realidad virtual de los medicamentos. *Química Viva*, 8(1), 25-34.

Meléndez-Herrada, E., Ramírez, M., Sánchez Dorantes, B. G., & Cervantes, E. (2010). Aportaciones de la genómica y la bioinformática al nuevo virus de la inflenza A (H1N1) y su impacto en la medicina. *Rev Fac Med UNAM*, 53(2), 76-82.

Mell, P., & Grance, T. (2011). The NIST defiition of cloud computing. U.S. Department of Commerce.

Meza, J., & Uribe, A. L. (2013). Implementación de dos nodos grid basados en clusters e integrados a grid Colombia a través de Renata, utilizando software libre. (Tesis de maestría) Universidad Autónoma de Manizales, Manizales, Colombia.

Rocha, M., Florentino, J., Corchado, E., Bustillo, A., Corchado, J (2009). *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*. Salamanca: Universidad de Salamanca.

Minetti, G. F. (2012). Problema de ensamblado de fragmentos de ADN resuelto mediante metaheurísticas y paralelismo. Paper presented at the XIV Workshop de Investigadores en Ciencias de la Computación. Tomado de: http://sedici.unlp.edu.ar/bitstream/handle/10915/19511/Documento_completo.pdf?sequence=1

Parisi, A., Parisi, F., & Díaz, D. (2006). Modelos de algoritmos genéticos y redes neuronales en la predicción de índices bursátiles asiáticos. *Cuadernos de economía*, 43(128), 251-284.

Patil, T. R., & Sherekar, S. (2013). Performance analysis of Naive Bayes and J48 classifiation algorithm for data classifiation. *International Journal of Computer Science and Applications*, 6(2), 256-261.

Pelta, D. A. (2013). *Algoritmos heurísticos en bioinformática*. (Tesis doctoral). Universidad de Granada: Granada, España.

Pham, D., & Karaboga, D. (2012). Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks: Springer Science & Business Media. Cardiff: University of Wales.

Pokkuluri Kiran Sree, I. R. B., SSSN Usha Devi .N. (2014). Cellular Automata and Its Applications in Bioinformatics: A Review. Global Perspectives on Artificial Intelligence (GPAI) Volume 2 (2) 16-22.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51-59.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, 202(4), 865-884.

Redondo-Gonzalez, E., de Castro, L. N., Moreno-Sierra, J., de las Casas, M. L. M., Vera-Gonzalez, V., Ferrari, D. G.,& Corchado, J. M. (2015). Bladder carcinoma data with clinical risk factors and molecular markers: a cluster analysis. *BioMed research international*. DOI: 10.1155/2738.

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Mineria de datos: Conceptos y tendencias. *Revista Iberoamericana de Inteligencia Artificial*, 10(29), 11-18.

Romero, R., Iglesias, E., & Borrajo, L. (2015). A Linear-RBF Multikernel SVM to Classify Big Text Corpora. *BioMed research international*. DOI: 10.1155/2738.

Schmidt, B. (2010). *Bioinformatics: High Performance Parallel Computer Architectures*: CRC Press.

Shamim, M. T. A., Anwaruddin, M., & Nagarajaram, H. A. (2007). Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, 23(24), 3320-3327.

Shinzawa, H., Jiang, J. H., Ritthiruangdej, P., & Ozaki, Y. (2006). Investigations of bagged kernel partial least squares (KPLS) and boosting KPLS with applications to near□infrared (NIR) spectra. *Journal of chemometrics*, 20(8□10), 436-444.

Tablada, C. J., & Torres, G. A. (2009). Redes Neuronales Artificiales. *Revista de Educación Matemática*, 24(3).

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1), 31-78.