3rd Workshop on MSc dissertations and PhD thesis in Artificial Intelligence (WTDIA'2006)
5th Best MSc dissertation / PhD thesis contest (CTDIA'2006)

CTDIA'2006 is the 5th best MSc Dissertation/ PhD Thesis contest in Artificial Intelligence. Its main goal is to award the three best academic work developed at universities from Ibero-American countries in the last two years.

WTDIA'2006 is the 3rd Workshop on Msc Dissertation and Phd Thesis in Artificial Intelligence. It provides an opportunity for Ph.D. and Ms.C. students to present their on going research with a panel of established researchers in artificial intelligence.

CTDIA'2006 and WTDIA'2006 are co-located with the International Joint Conference SBIA/IBERAMIA /SBRN'2006.

For CTDIA'2006 we had 30 submissions (10 Ph.D. Thesis and 20 M.Sc. Dissertations). The contest was carried out in two phases. In the first phase, all papers were peer-reviewed by three referees of the area and 15 were selected for the second phase of evaluation (5 Ph.D. Thesis and 10 M.Sc Dissertations). In the second phase, two Award Program Committees (APC), one for Ph.D. Thesis and the other for M.Sc Dissertations, were responsible for selecting the three best Thesis/Dissertation in each category. To the fairness of this process, in each category, each member of the APC re-evaluated all papers selected in the first phase.

For WTDIA'2006 we had 39 papers (14 Ph.D. Thesis and 25 M.Sc. Dissertations). All papers were peer-reviewed by at least two referees of the area. As a result of this selection process, 22 papers (11 Ph.D. Thesis and 11 M.Sc. Dissertations) were accepted for presentation at the Workshop.

The CTDIA'2006 and WTDIA'2006 chairs would like to thank the authors for submitting their work and the Award and Program Committee members, as well as the other reviewers, for their dedication in the reviewing process.

Chairs:

José Augusto Baranauskas (FFCLRP/USP, Brazil)
Maria Carolina Monard (ICMC/USP, Brazil)

Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 5th Best MSc dissertation/PhD thesis contest (CTDIA'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Programación Lógica Inductiva en la exploración de regularidades en el Cromosoma 2 del *Plasmodium Falciparum*

Autor: José H. López Prato.
Afiliación: Laboratorio de Computación de Alto Rendimiento, Universidad del Táchira, San Cristóbal, Táchira, Venezuela. jlopez@unet.edu.ve

Supervisor: Jacinto Dávila
Afiliación: CESIMO: Centro de Simulación y Modelos, Fac. Ingeniería. Universidad de Los Andes, Mérida, Venezuela, jacinto@ula.ve.

Biologo asesor: Alejandro Mujica
Afiliación: CeCALCULA: Centro Nacional de Cálculo Científico, Parque Tecnológico. Mérida, Venezuela. alejo@ula.ve.

Resumen

Se describe el desarrollo de experimentos de aprendizaje automático realizados en el cromosoma 2 del parásito de la Malaria Plasmodium falciparum. El objetivo ha sido describir sitios de splicing GT y AG. Se proponen cuatro etapas para la elaboración de un predictor de genes basado en reglas, de las cuales dos se presentan aqui. Se presenta la programación lógica inductiva (ILP) y se describen modelos de experimentos. En la primera etapa se aprenden y validan reglas sobre posibles secuencias conservadas (PSC) alrededor de los sitios de splicing; además de las zonas de ramificación y ricas en pirimidinas. La segunda etapa hace uso de las reglas provenientes de la primera e incluye otros conceptos y restricciones reportados en otros trabajos, resultando éstos validados o descartados en el proceso. Se analizan los resultados obtenidos mediante tablas de contingencia apoyadas en el estadístico Chi-cuadrado. En la primera etapa, las PSC y la zona rica en pirimidinas se muestran como los criterios que mejor describen los sitios de splicing, descartándose el consenso reportado para la zona de ramificación. En la segunda etapa se generalizan reglas para los sitios de splicing, incluyendo otros conceptos y restricciones. En esta etapa destacan las PSC, las restricciones y las transiciones de contenido GC y AT, como mejores descriptores de sitios de splicing. Las tablas de contingencia sugieren que las reglas pueden ser efectivas para identificar tales sitios y constituyen evidencia de la factibilidad de un predictor genético basado en ILP. Los scripts ILP desarrollados y los conceptos de soporte (Knowledge Background) asociados, han sido liberadas bajo licencia pública GNU.

Palabras clave: Malaria, *Plasmodium falciparum*, Bioinformática, Aprendizaje Automático, *Splicing*, ILP.

Nivel: Disertación de Maestria. **Fecha de conclusión**: Junio 2004.
Scripts ILP disponibles en: http://sourceforge.net/projects/simulants
Tesis a texto completo: ftp://lcar.unet.edu.ve/Papers

1. Introducción

La dificultad asociada a la definición de la estructura de un gen y su posible funcionalidad está relacionada con varios aspectos. Los genes eucariotas codificantes de proteínas están divididos en Introncs y Exones (ver Fig. 1). Para identificar éste tipo de genes es esencial identificar los límites exon/intron. Existen señales que marcan sistemáticamente la ubicación de estos bloques en la secuencia. Sin embargo, las mismas señales se **repiten** con frecuencia y en otras ubicaciones en donde no existen evidencias experimentales de su funcionalidad (falsos positivos). Por otro lado algunas señales con funcionalidad comprobada se escapan de los predictores actuales (falsos negativos). Esto indica que existen falsos inicios o falsas terminaciones que dificultan la predicción de la estructura de los genes. Se han desarrollado métodos más elaborados que el simple rastreo de esas "marcas" para establecer la estructura real de un gen (Padgett & Burge, 2005; Salzberg et al, 1998). Existen predictores basados en modelos de Markov que proponen la estructura general de genes eucariotas (Majoros et al, 2004; Pertea et al, 2001; Burge, 1998). Otras aproximaciones basadas en estudios de correlación, detección de señales y alineamientos, generan modelos para zonas específicas (e.g. sitios de splicing o exones) que luego pueden ser empleados como submodelos de predictores mas generales (Hsieh et al 2005; Wang et al 2004; Arita et al 2002). A diferencia de esos trabajos, nuestro enfoque propone la construcción de descripciones estructurales que se descubran y originen desde las secuencias de ADN, usando procesos inductivos para la generación de reglas. Nuestra meta es generalizar desde las

secuencias las evidencias o criterios que en la actualidad se emplean para validar o corregir estructuras genéticas propuestas *in silico* o *in vivo*. Nuestra metodología propone una aproximación *bottom-up* para reconstruir tales evidencias y experimentar posibles generalizaciones y ajustes de las mismas.

A grosso modo se puede decir que exones e intrones son transcritos a ARN, pero en el procesamiento a ARN mensajero maduro (mARN), los intrones son extraídos por "splicing" (Padgett & Burge, 2005; Wen-Hsiung, 1999; Lehninger, 1999; Rawn, 1989). Un marco de lectura contenido en los exones, en general el más extenso posible, lleva el código necesario para la síntesis de una proteína determinada. Por su naturaleza generalmente codificante, el tamaño y secuencia de exones homólogos tiende a conservarse evolutivamente, no así los intrones, cuyo tamaño y secuencia pueden variar grandemente por ser más tolerables a la acumulación de mutaciones y secuencias transposables. Sin embargo, las señales de splicing que determinan los límites exon/intron y que guían el proceso mismo de eliminación de intrones y concatenación de exones, están contenidas en la secuencia de todo gen, en particular dentro de los intrones y son comunes para todos los genes. De tal modo que Puede plantearse la siguiente Regla/Plantilla para caracterizar las triadas exon/intron/exon:

exon/GT-intron-AG/exon **(A)**

A su vez, el intrón puede detallarse del siguiente modo:

5'-AG/GTAAGT---intron----YNCTRAC------YnNAG/G-3' **(B)**

Donde "/" marca los límites exon/intron, Y es alguna pirimidina (T o C), Yn es un fragmento de ~10 pirimidinas, R es una purina (A o G). La A presente en la subcadena RAC (ver regla B) es una Adenina en algún lugar de la parte interna del intron, esencial para la reacción intermediaria de separación del intron. N es cualquier base. Toda la secuencia entre las dos barras inclinadas ("/ /") es un intron y el tamaño es muy variable. El problema tratado en este trabajo puede precisarse de la siguiente manera: El conjunto de reglas que rigen el proceso de splicing del ARN primario, se encuentra en la secuencia de ADN genómico y se aplica de manera dispersa a lo largo de cada gen. Tal conjunto de reglas está parcialmente definido y se compone de conocimiento inferido mediante métodos estadísticos validado (en algunos casos) experimentalmente.

2. Trabajos relacionados

El presente trabajo propone un modo de estudio alternativo basado en aprendizaje automático (Ian y Eibe, 2005; Mitchell, 1997) y el método empleado es la programación lógica inductiva o Inductive logic Programming ILP (Muggleton y Firth 2003; Muggleton, 1995). La programación ILP ha sido empleada en diversos problemas relacionados con la bioinformática que incluyen predicción de estructuras secundarias de proteínas, predicción de carcinogenicidad y sistemas metabólicos (Tamaddoni et al, 2006; Muggleton, 2005; Turcotte et al, 2001; Turcotte et al, 1998; Finn et al, 1998; Srinivasan et al, 1997). El análisis de regularidades en secuencias de ADN empleando reglas lógicas o ILP ha sido reportado en otros trabajos. Kolchanot et al (2002) presentan un sistema que usa reglas lógicas ponderadas probabilísticamente para describir promotores y discriminar la funcionalidad de los genes asociados. Wren et al (2005) reportan el uso de árboles de decisión para descubrir y organizar reglas acerca de características compartidas entre genes, exones, *repeats* e islas CpG, de un mismo genoma o entre genomas diversos. King et al (2004) emplean ILP y árboles de decision para predecir la funcionalidad asociable a marcos de lectura. Badea (2003) emplea ILP para discriminar la funcionalidad de genes estudiados mediante *microarrays*, expresando las reglas inherentes en términos de una ontología. Keles et al (2004) analizan mediante regresión lógica correlaciones entre los sitios de enlace en promotores asociados a factores de transcripcion y sus efectos en la regulacion de los genes correspondientes. Estos trabajos indican que ILP ha sido y está siendo empleada en el análisis de secuencias. Sin embargo la generalización de reglas para proponer la estructura de un gen por medio de ILP parece no haber sido abordada aún. Lo mas cercano, después de una revisión detallada, es la predicción de la funcionalidad asociable a marcos de lectura o genes. Hasta ahora parece no existir un trabajo basado en ILP, organizado según los objetivos y la metodología que aquí presentamos.

3. Motivación

El presente trabajo tiene como objetivo general el desarrollar procesos de aprendizaje automático por capas mediante ILP, que validen, cuestionen y traten de ajustar las reglas que describen sitios de splicing. Nuestro trabajo

debe conducir, en posteriores etapas, a la construccion de una base de conocimiento capaz de identificar desde regularidades basicas hasta posibles estructuras genéticas en secuencias de ADN.

4. Metodo de investigación y Enfoque

Nuestro trabajo está dirigido a la construcción de experimentos de aprendizaje automático organizados por niveles para producir representaciones de conocimiento que puedan ser jerarquizadas y manipuladas por biólogos. El tipo de reglas descubiertas deben describir desde regularidades básicas tales como posibles secuencias conservadas o regularidades aún más generales para la descripción de sitios de splicing. Tales reglas son minadas y evaluadas empleando motores de inferencia provenientes de la programación logica inductiva, cuya eficiencia y confiabilidad ha sido probada previamente. Nuestra metodología *ad hoc* procura el desarrollo de niveles de aprendizaje organizados por capas para descubrir e interrelacionar las características estructurales de cada de uno de los conceptos presentados en la tabla I. Debe entenderse que cada nivel recibe conceptos aprendidos en el nivel inmediatamente anterior. Para lograr lo anterior se organizan experimentos de aprendizaje para cada capa empleando criterios biológicos bien establecidos, acerca de las características que generalmente están asociadas a cada uno los conceptos listados. Tales criterios permiten la definición de una base de conocimiento inicial descrita mediante reglas lógicas. Este documento describe los primeros resultados logrados para las dos primeras capas de aprendizaje.

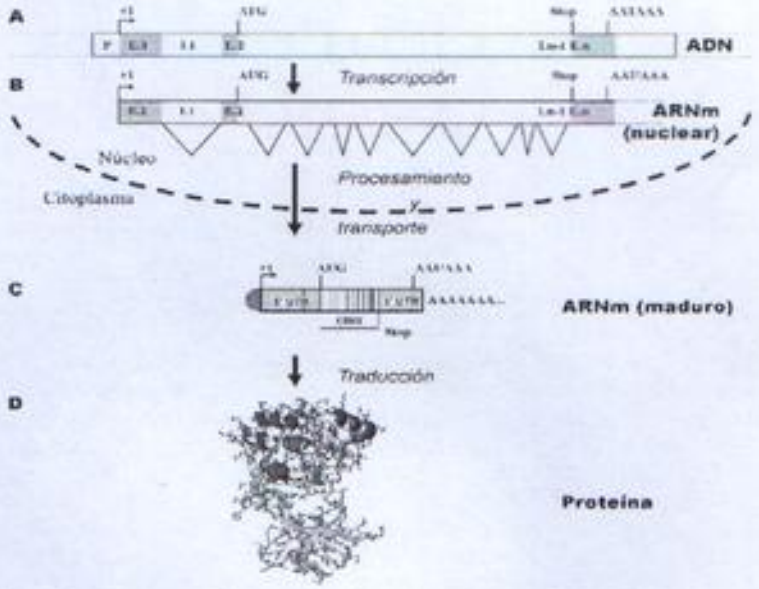

Fig 1. Estructura general y expresión de un gen eucariota codificante de Proteínas.

A) Una región de ADN genómico dada, de longitud indeterminada (líneas punteadas), contiene un gen con un número N de exones (E.1,...,E.n) separados por un número N-1 de intrones (I.1,..I.n-1). B) La región promotora P determina el inicio de transcripción (+1) que resulta en el ARN mensajero nuclear (o primario) con exones e intrones. B->C) El ARN es procesado: los intrones son escindidos (splicing) representado aquí por líneas diagonales, en el extremo 5' se añade la estructura CAP (semicírculo) y una señal de poliadenilación (AAUAAA) determina un corte a unas 25 bases aguas abajo y la adición de una cola de adeninas (AAAAAA...). El ARN mensajero maduro (ARNm) es transportado al citoplasma. Flanqueado por las regiones no traducibles (UTR) y delimitado por las señales de inicio (AUG) y parada (UAA, UGA o UAG), se encuentra un marco abierto de lectura (ORF) que codifica la información para la síntesis (traducción) de una proteína (D). Las señales de inicio de lectura, de parada, de poliadenilación así como los límites exón/intrón son comunes para todos los genes, no así la información contenida en el ORF, el inicio de transcripción, el número de exones e intrones ni su tamaño.

4.1 Tipo de criterios a-priori, empleados para la caracterización de sitios de splicing

Nuestros ejemplos y contraejemplos se han extraído de la reanotación del cromosoma 2 del *P. falciparum* realizada por (Huestis y Fisher, 2001). A continuación algunos de los criterios empleados para el estudio (Huestis y Fisher, 2001; Wen-Hsiung, 1999; Lehninger, 1999; Rawn, 1989).

1. Algunos nucleótidos suelen mostrar cierta distribución en el interior del intron. Las Aes tienden a estar hacia el extremo 5´ del intron, las T´s hacia el extremo 3´, y la zona poly-RY hacia el centro.
2. La repetición de dinucleótidos es típica en intrones. Todas las formas [RY]n con n>5 generalmente se asignan a intrones.
3. Cuando ocurre un sitio de splicing GTGT éste suele asociarse al primer dinucleótido. Esto se decide así puesto que es extraño que un exon finalice con GT.
4. La zona de enlace o ramificación es una zona bastante cargada de T´s que suele ser interrumpida por A, C o G y que suele ubicarse a unas 40 b "aguas arriba" del sitio de splicing AG.
5. Una región rica en bloques AT precedida de una señal GT es considerada parte interna de un intron. Las zonas GT [AT]n generalmente ocurren bastante después del sitio de splicing GT, dado que la zona [AT]n ocurre hacia el centro del Intron y no al principio.
6. Una señal GT seguida de A o AA es un sitio de splicing bastante probable. GTC al contrario, suele ser un sitio de splicing falso. Los sitios de splicing GTAA y GTA ocurren aproximadamente dos terceras partes de las veces en que se detecta una señal GT válida.
7. Cuando hay varias señales GT agrupadas se da preferencia a aquellas precedidas por G o A, dado que son sitios de splicing más probables; sin embargo, tal decisión no debe sacrificar zonas de ADN con marcada tendencia GC.
8. Una señal AG verdadera suele estar precedida por una citosina, una base cualquiera y el inicio de una la zona rica en pirimidinas seguida de la zona de ramificación. El sitio candidato AG suele estar seguido de una guanina.
9. Un sitio de splicing AG es poco probable cuando tiene otro sitio similar cercano aguas arriba.
10. El splicing alternativo se reconoce cuando hay más de una manera de escindir un intron. Esto implica la presencia de más de una señal GT o AG positivas.
11. Las zonas codificadoras suelen ser más ricas en GC que las no codificantes.
12. Las zonas codificantes suelen tener codones particulares que suelen repetirse. En particular GAA, GAT, AAT.

Nivel de aprendizaje	***Conceptos a generalizar e interrelacionar***
I	Posibles secuencias conservadas, zona de ramificación, zona Rica en pirimidinas, recurrencia de *repeats*, zona de poliadenilación, transición de contenido AT y GC, Otros.
II	Sitios de inicio, sitios de *splicing*, sitios de parada, sitios de splicing alternativo, sitios de inicio alternativos.
III	Intrones, Exones
IV	Genes

Table I._ Niveles de aprendizaje automatico para descubrir modelos genómicos.

4.2 Uso de un motor de inferencia inductiva. Progol.

Progol es un sistema inductor de reglas que opera sobre el espacio de modelos, o reticulado de hipótesis, asociado a una teoria lógica (Hogger, 1990). El sistema toma como entrada una teoria con el conocimiento establecido **B**, y los ejemplos **E**⁺, y contraejemplos **E**⁻, pertinentes al concepto que se desea generalizar. En general, cualquier experimento de aprendizaje basado en Progol, requiere que se definan los siguientes componentes: Teoria parcial (o conceptos); Definiciones estructurales (modos y tipos); Restricciones y Podas; Ejemplos y Contraejemplos. La herramienta emplea tres etapas para realizar sus procesos de aprendizaje 1) Construcción de la cláusula más específica, 2) Construcción y recorrido del reticulado de hipótesis y 3) Ejecución del algoritmo de cubrimiento (Muggleton, 1995). A continuación una breve descripción del modo en que Progol se emplea en el presente trabajo.

4.2.1 Construcción de la cláusula más específica

Para explicar esta primera etapa se emplea el concepto *Posibles secuencias conservadas "aguas arriba" de una senal de splicing GT*, *pscgtup*(`Bases, Señal, Dirección`), encargado de definir posibles secuencias conservadas (PSC) a la izquierda de sitios GT. Tal concepto debe ser capaz de recibir una secuencia de `Bases`, la `Señal` a buscar y una `Dirección` de estudio. La idea es generalizar reglas para el concepto tales que, dados los argumentos anteriores, se pueda responder si la secuencia posee o no un sitio de splicing GT. Para generalizar el concepto *pscgtup*, Progol emplea reiteradamente el concepto `b(Base, Zona, Resto)`, responsable de determinar el tipo de `Base` al inicio de una `Zona` de estudio y devolver el `Resto` de la misma. El primer paso del