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 Mujíca
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 aquí. 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 Introns 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 ontologia. 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 aqui 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.

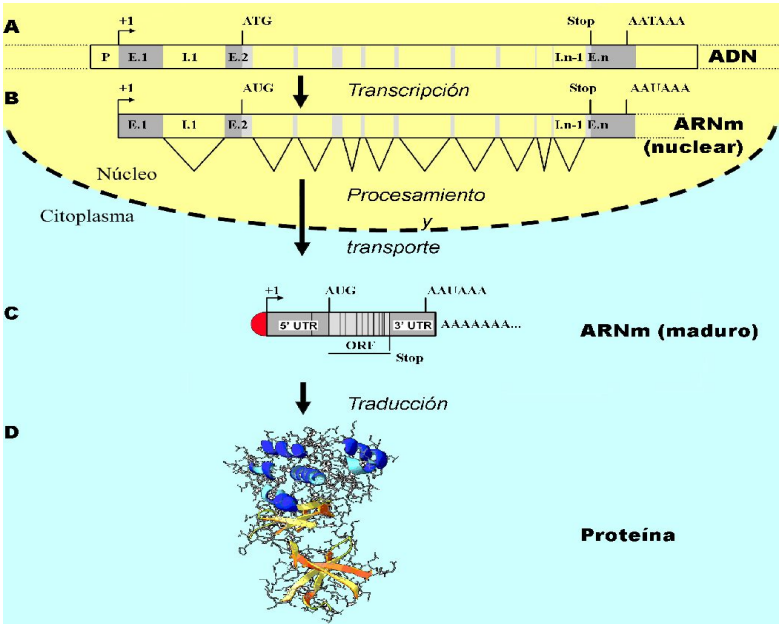

Fig 1. Estructura general y expresión de un gen eucariota codificante de Proteínas.

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. Flanquedo 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 teoría lógica (Hogger, 1990). El sistema toma como entrada una teoría con el conocimiento establecido **B, y** los ejemplos $\mathbf{E^+}$**,** y contraejemplos $\mathbf{E^-}$**,** pertinentes al concepto que se desea generalizar. En general, cualquier experimento de aprendizaje basado en Progol, requiere que se definan los siguientes componentes: Teoría 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

inductor es definir una regla candidata inicial, siendo ésta conocida como la cláusula más específica MSC (Most Specific Clause) (Muggleton, 1995). Para construir la primera regla Progol toma uno de los ejemplos positivos suministrados y define la zona de la secuencia para la que se generalizarán reglas. Para ello Progol emplea el concepto `zona_es(Dirección, Bases, Señal, Posición, Zona).` Este concepto recibe la secuencia de `Bases` que será sujeto de análisis, la `Dirección` de estudio y la `Señal` que se debe emplear para delimitar el sitio a estudiar; además el concepto devuelve la `Posición` en la que se encuentra la `Señal` y la `Zona` para la que deben generalizarse reglas. Los conceptos `zona_es/5` y `b/3,` se aplican empleando DCGs (Definite Clause Grammars), técnica de procesamiento proveniente de la lingüística computacional (Siu-wai, 2001). La regla C, presentada a continuación, muestra el tipo de reglas que Progol logra establecer.

```
pscgtup(Sequencia,Senal,Direccion) si
            zona_es(Direccion,Sequencia,Senal,Posicion,Zona) y
             b(r,Zone,Rest_Zona) y b(a,Rest_Zona,Nuevo_Rest_Zona).
```

(C)

pscgtup: Posibles secuencias conservadas "aguas arriba" de una senal de splicing GT

Interpretacion: Una **Secuencia** de bases incluye un sitio de splicing GT *si*, en una delimitada **Zona** que incluye una senal GT, una base purina **r** es seguida por una base adenina a "aguas arriba"respecto de la senal GT.

4.2.2 Construcción y recorrido del reticulado de hipótesis

La cláusula más específica debe ser optimizada. La finalidad es determinar si existe una manera más eficiente de organizar los conceptos de la regla. Para ello se recorre un reticulado de reglas alternativas (vistas como subconjuntos de la primera). El recorrido del reticulado permite eliminar la presencia de cualquier concepto redundante, establecer un mejor orden de los conceptos y evaluar la calidad de las alternativas propuestas en la medida en que se van definiendo éstas. Para cada hipótesis se chequean las restricciones establecidas (si acaso existen). Si alguna hipótesis satisface alguna restricción, entonces el sistema regresa hasta el punto en el que la restricción deja de ser válida e incorpora otro componente de la MSC, creando una nueva regla alternativa para explorar.

4.2.3 Aplicación del algoritmo de cubrimiento

Este algoritmo se encarga de evaluar la totalidad de los ejemplos positivos y determina cuáles ejemplos son deducibles mediante una hipótesis recién propuesta por el inductor. Cuando una hipótesis califica como regla, cada ejemplo que la satisface es extraído del conjunto inicial de ejemplos. En resumen: Tomado el primer ejemplo, primero se define una MSC a partir de éste, partiendo de la MSC se recorre un reticulado de hipótesis, se evalúa cada una de ellas para hallar la mejor y se determinan los ejemplos positivos y negativos que son cubiertos por la misma. Una vez determinada la mejor hipótesis, los ejemplos son extraídos del total de ejemplos positivos suministrados. Progol, procede, de allí en adelante, a tomar el primer ejemplo positivo disponible y repite todo el proceso nuevamente. Esto se repite hasta que hayan sido generadas todas las reglas necesarias para implicar el total de los ejemplos positivos, procurando siempre cubrir la menor cantidad de ejemplos negativos (en caso de que se le permita tal flexibilidad).

4.2.4 Presentación de resultados

La tabla I muestra el resultado final en el análisis de PSC en sitios GT para el cromosoma 2 del *P. falciparum*. Las reglas halladas fueron evaluadas por el mismo inductor mediante un estudio estadístico, basado en tablas de contingencia, que comparan predicciones (apoyadas en las reglas) con una clasificación aleatoria basada en una distribución probabilística uniforme. Los valores Chi-square = `24.64` y Chi-square probability = `0.0000`, estiman que el resultado está muy distante a una simple clasificación azarosa. Puede observarse en la columna A de la tabla de contingencia que se manejaron 39 ejemplos positivos mientras que la columna ~A muestra un total de 36 ejemplos negativos. Las filas P y ~P especifican cuántos del total de ejemplos son clasificados como positivos y negativos. Puede verse que las reglas logran clasificar adecuadamente a 33 de los 39 ejemplos positivos y agregan allí (erróneamente) 9 de los 36 negativos suministrados. También puede notarse que las reglas, desafortunadamente, no logran clasificar a 6 de los positivos, mientras que 27 de los negativos son correctamente clasificados. Los ejemplos evaluados corresponden aproximadamente a un 30% de los sitios GT verdaderos reportados en (Huestis y Fisher, 2001). El 70% restante se empleó en el proceso de aprendizaje de reglas. El inductor presenta estadísticos que permiten ponderar la precisión de las reglas. Estos son los valores C: 83, 98, 24, 0 `(ver tabla)` que pueden reescribirse como `C: f, p, n, h` siendo 'f' el parámetro que mide lo precisa que es la regla. 'f' es una medida del poder predictivo y de compresión de cada regla hallada. Los parámetros `p` y `n` corresponden al número de ejemplos positivos y negativos cubiertos por la regla mientras que `h` indica el número de conceptos que aún deben probarse

para considerar completa una regla. La Chi cuadrado mide cuán aleatorio es el conjunto de resultados predichos por las reglas que se someten a evaluación (estadístico Chi-square) y qué tan probable es que esos resultados equivalgan a selecciones al azar (estadístico Chi-square probability).

5 Experimentos

Nuestro trabajo esta organizado por etapas de aprendizaje. Las dos primeras etapas se describen a continuación. Para cada experimento se organizaron archivos de entrenamiento y archivos de evaluación. Los primeros se obtuvieron de las anotaciones del cromosoma 2 del *P. falciparum*, suministradas por (Huestis y Fisher, 2001). Para el caso de ejemplos negativos se optó por dos aproximaciones: La primera, generar ejemplos tomando señales falsas detectadas en zonas intergénicas. Para la segunda, ubicada una señal ATG verdadera, por ejemplo, se programaron *scripts* que desplazan la ubicación del sitio de inicio ya sea hacia la derecha o hacia la izquierda. Los ejemplos y contraejemplos se dividieron en dos conjuntos de datos destinados a entrenamiento y evaluación, tomando 70% y 30% de los casos disponibles para cada fin.

5.1 Primera etapa de aprendizaje

5.1.1Estudio de posibles secuencias conservadas

La organización de los experimentos relativos a este concepto se explica en la sección 4.2.1.

5.1.2 Zona de ramificación y zona rica en pirimidinas.

Para descubrir la posicion de la zona de ramificación se empleó el consenso YYRAY. El concepto para la zona de ramificación es `zona_ramifi/3` que usa el concepto *`ramificación/4,`* diseñado para ubicar el consenso y devolver la posición en la que se encuentra (/n para indicar el número de argumentos). Se emplean además los conceptos *`entre/2`* y *`y/2`*, para aprender el rango de posiciones en el que generalmente se ubica tal zona. La estrategia es similar para la zona rica en pirimidinas. La tabla III resume los resultados obtenidos en el análisis de zonas.

5.2 Segunda etapa de aprendizaje

En esta etapa se incorporan varios de los criterios para caracterizar los sitios de splicing (GT/AG), expuestos en la sección 4.1; por ejemplo, el uso de las transiciones de zonas no codificantes a zonas que si lo son. También existen reglas que permiten descartar posibles sitios de splicing; por ejemplo, está bien establecido que aquellos sitios de splicing GT precedidos de otros sitios cercanos suelen ser pobres candidatos (sección 4.1, ítem 5). Para manejar los criterios anteriores se programaron conceptos como: *`trans_AT/4`* y *`trans_GC/4`* que determinan si hay transiciones importantes de contenido AT o GC en un sitio determinado; *`psc_gt/4`*, que establece si un sitio determinado satisface o no la distribución de posibles secuencias conservadas previamente aprendida. También se incorpora el concepto *`zona_rica_pirimidinas/4`*, aprendido en la primera etapa. Además de tales criterios, la segunda etapa incorpora restricciones del tipo: *`no_gt_vec_prec_ag/2`*, que valida que el sitio GT candidato no presente otro sitio GT vecino precedido por A o G; *`no_gt_prec_gt/2`*, que *d*etermina si el sitio GT en estudio no está precedido por otra señal GT y `no_ag_vecino/2` , que determina si el sitio AG en estudio no está precedido por otra señal AG vecina. La tabla IV resume los resultados obtenidos durante esta etapa.

6 Resultados

6.1._ Primera etapa

6.1.1 Posibles sitios conservados GT

La tabla I muestra un resultado que concuerda con lo reportado en sección 4.1, ítem 6 y 7. La primera regla para el análisis de PSC a la derecha de sitios GT valida lo reportado en el ítem 6, indicando además que en efecto se trata de la regla más efectiva. Se postula una regla adicional menos efectiva pero con diferencias importantes respecto de la primera. Se concuerda en que el sitio de splicing suele estar seguido por una adenina pero la regla postula adicionalmente los nucleótidos ACG y ATG como característicos a la derecha del sitio de splicing GT. Esto último no se encuentra reportado en el trabajo de (Huestis y Fisher, 2001) y podría constituir una ampliación de las secuencias consenso hasta ahora conocidas. El ítem 7 de la sección 4.1 señala una regularidad hacia la izquierda que solo es válida cuando se agrupan señales GT. Nuestro resultado indica que la misma es cierta en general, proponiendo GA o AA como dinucleótidos recurrentes, lo que amplía el criterio de clasificación. Los valores del test de Chi-cuadrado correspondiente al análisis de PSC en sitios de splicing GT indican que la regla es estadísticamente efectiva clasificando los ejemplos del concepto.

Aguas arriba:
C:83,98,24,0 pscgtup(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(r,Zona,Rest_Zona) *y* b(a,Rest_Zona,Nuevo_Rest_Zona).

Aguas abajo:

Regla 1:
C:89,72,12,0 pscgtdown(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(a,Zone,Rest_Zona) *y* b(**a**,Rest_Zona,Nuevo_Rest_Zona).

Regla 2:
C:16,8,1,0 pscgtdown(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(a,Zone,Rest_Zona) *y* b(**y**,Rest_Zona,Nuevo_Rest_Zona),
b(a,Nuevo_Rest_Zona,Ultimo_Rest_Zona).

Contingency table (pcsgtup/3)=

	A	~A	
P	33	9	42
	(21.8)	(20.2)	
~P	6	27	33
	(17.2)	(15.8)	
	39	36	75

Contingency table (pcsgtdown /3)=

	A	~A	
P	32	14	46
	(17.2)	(28.8)	
~P	8	53	61
	(22.8)	(38.2)	
	40	67	107

Overall accuracy	Chi-square	Without Yates correction	Chi-square probability
80.00% +/- 4.62%	24.64	27.00	0.0000
79.44% +/- 3.91%	33.33	35.70	0.0000

Tabla I._ Posibles secuencias conservadas en sitios de spling GT

Aguas arriba:
C:54,57,12,0 pscagup(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(**y**,Zone,Rest_Zona) *y* b(**n**,Rest_Zona,Nuevo_Rest_Zona),
b(**y**,Nuevo_Rest_Zona,Ultimo_Rest_Zona).

Aguas abajo:

Regla 1:
C:66,18,3,0 pscagdown(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(**g**,**Zone**,Rest_Zona) *y* b(**t**,Rest_Zona,Nuevo_Rest_Zona).

Regla 2:
C:1,6,1,0 pscagdown(Sequencia,Senal,Direccion) *si*
zona_es(Direccion,Sequencia,Senal,Posicion,Zona) *y*
b(**r**,Zone,Rest_Zona) *y* b(**g**,Rest_Zona,Nuevo_Rest_Zona).

Contingency table (cagi/3)=

	A	~A	
P	36	7	43
	(21.5)	(21.5)	
~P	4	33	37
	(18.5)	(18.5)	
	40	40	80

Contingency table (cagd/3) =

	A	~A	
P	8	0	8
	(4.0)	(4.0)	
~P	32	40	72
	(36.0)	(36.0)	
	40	40	80

Overall accuracy	Chi-square	Without Yates correction	Chi-square probability
86.25% +/- 3.85%	39.42	42.29	0.0000
60.00% +/- 5.48%	6.81	8.89	0.0091

Tabla II._ Posibles secuencias conservadas en sitios de spling AG

6.1.2 Posibles sitios conservados AG

La tabla II muestra que la primera regla concuerda con lo reportado en sección 4.1, ítem 8. La primera regla es más general que lo reportando en (Huestis y Fisher, 2001) ya que propone una pirimidina seguida de una base cualquiera y una pirimidina. El grupo de reglas propone una guanina inmediata hacia la derecha de la señal AG, coincidiendo con lo indicado en sección 4.1, ítem 8. Debe observarse que también se propone una timina como segundo nucleótido conservado. Esto no esta reportado en ninguna de las referencias consultadas. Puede observarse que los valores Chi cuadrado indican mayor verosimilitud para la regla que describe patrones conservados a la izquierda del sitio de splicing mientras que las otras dos son menos confiables. La tabla de contingencia para el predicado `pscagdown/3` indica que las reglas son buenas discriminadoras de verdaderos negativos pero no verdaderos positivos.

6.1.3 Zonas de ramificación y zona rica en pirimidinas.

El resultado indica que el consenso para la zona de ramificación no es confiable, lo que en efecto refuta el consenso propuesto. Debe decirse sin embargo que el motor descubre una banda de posiciones para la tal zona que se corresponde con lo reportado (ver sección 4.1, ítem 4). Los resultados pueden observarse en la tabla III. Se considera necesario un mejor estudio basado en experimentos de alineamiento, por ejemplo, para caracterizar mejores consensos. En el caso de la zona rica en pirimidinas se obtuvo un mejor resultado, validándose la presencia de ésta en las proximidades de los sitios de splicing AG (ver sección 4.1, ítem 8). Los valores de Chi Cuadrado muestran a este consenso como un criterio de decisión muy importante.

6.2 Segunda etapa

6.2.1 Sitio GT

En este caso se señalan la transición de contenidos AT y la distribución de PSC a la izquierda de los sitios GT como las reglas discriminatorias más sólidas. En la primera regla se esperaría una transición del tipo codificante a no codificante, sin embargo la transición descubierta es del tipo `'no_cod->cod'`. Lo que la primera regla establece en este caso es que al final del exon el contenido AT es mas bajo que el correspondiente al inicio del intron. Se aclara que el estudio de las proporciones se hace considerando solo varias decenas de bases en cada dirección. El valor de Chi cuadrado es bastante alto. Sugiere que el par de conceptos presentados en las reglas tienen un peso importante en la identificación de sitios de splicing GT.

6.2.2 Sitio AG

Pueden observarse tres reglas para identificar el sitio de splicing AG. Primero, se establece una zona rica en pirimidinas y la ausencia de otra señal AG vecina aguas arriba. En segundo lugar, se observa que es usual detectar transiciones GC `'no_cod->cod'`. Por último, se establece la presencia de transiciones AT alto a bajo (`'cod->no_cod'`), en conjunción con la distribución de conservados a la derecha del sitio AG aprendida en la primera etapa. El valor Chi cuadrado y la probabilidad correspondiente definen un nivel alto de confianza en el resultado.

```
zona_ramifi(Sequencia,Senal,Direccion) si
          ramificacion(Direccion,Sequencia,Senal,Posicion) y entre(Posicion,35) y y(Posicion,40).

 zona_rica_pirimidinas(Sequencia,Senal,Direccion) si
          pirimidinas(Direccion,Sequencia,Senal,Posicion) y  entre(Posicion,2) y y(Posicion,7).

 zona_rica_pirimidinas(Sequencia,Senal,Direccion) si
          pirimidinas(Direccion,Sequencia,Senal,Posicion) y  entre(Posicion,0) y y(Posicion,12).
```

Contingency table (zona_ramifi/3)=

	A	~A	
P	1 (1.1)	1 (0.9)	2
~P	34 (33.9)	29 (29.1)	63
	35	30	65

Contingency table (zona_rica_pirimidinas/3)=

	A	~A	
P	28 (17.8)	5 (15.2)	33
~P	7 (17.2)	25 (14.8)	32
	35	30	65

Overall accuracy	Chi-square	Without Yates correction	Chi-square probability
46.15% +/- 6.18%	0.37	0.01	0.5422
81.54% +/- 4.81%	23.45	25.92	0.0000

Tabla III._ Resultados análisis zona de ramificación y zona rica en pirimidinas.

```
sitio_splicing_gt(Sequencia,Senal, Posicion) si trans_AT(Sequencia,Posicion,Senal,'no_cod->cod').

sitio_splicing_gt(Sequencia,Senal, Posicion) si pcsgt_gt(Sequencia,Posicion,Senal,izq).

sitio_splicing_ag(Sequencia,Senal, Posicion) si

        zona_rica_pirimidinas(Sequencia,Posicion,Senal,izq) y no_ag_vecino(Sequencia,Posicion,Senal).

sitio_splicing_ag(Sequencia,Senal, Posicion) si trans_GC(Sequencia,Posicion,Senal,'no_cod->cod').

sitio_splicing_ag(Sequencia,Senal, Posicion) si

        trans_AT(Sequencia,Posicion,Senal,'cod->no_cod'),conservados_ag(Sequencia,Posicion,Senal,der).
```

Contingency table (sitio_splicing_gt/2)=

	A	~A	
P	35 (18.1)	0 (16.9)	35
~P	10 (26.9)	42 (25.1)	52
	45	42	87

Contingency table (sitio_splicing_ag/2)=

	A	~A	
P	31 (18.3)	3 (15.7)	34
~P	4 (16.7)	27 (14.3)	31
	35	30	65

Overall accuracy	Chi-square	Without Yates correction	Chi-square probability
88.51% +/- 3.42%	51.47	54.65	0.0000
89.23% +/- 3.84%	36.89	39.98	0.0000

Tabla IV._ Resultados generalización conceptos sitios de splicing GT y AG.

Conclusiones

Hemos creado una plataforma computacional, basada en el inductor Progol, para identificación de estructuras genéticas, a partir de datos publicados del *P. falciparum*. Dada una descripción formal de ejemplos positivos y negativos y una base de conocimiento inicial, la plataforma produce reglas lógicas que describen regularidades estructurales. La primera etapa de aprendizaje se ha centrado en el estudio de PSC, zonas ricas en pirimidinas, y de ramificación. Los dos primeros definen criterios de decisión importantes mientras que el último resulta muy pobre en la identificación de los sitios de interés. Sin embargo el motor de inferencia propone una posible banda para su ubicación. El análisis de PSC debe observarse con atención. Se logran validar varias de las regularidades conocidas

pero acompañadas de otras no reportadas. Cuando esto ocurre en una misma regla lógica las regularidades resultan correlacionadas, debido a que son elementos en conjunción constitutivos de una misma regla. Esto debe cotejarse con cuidado con lo reportado en otros trabajos, en particular los referentes a eucariotas superiores. En otros casos, se validan los consensos reportados pero existen simultáneamente otras reglas que reportan otros patrones de regularidad más eficientes según se lee en el estadístico f asociado a cada regla.

La segunda etapa de aprendizaje tomó los resultados obtenidos en la primera e incorporó otros conceptos de interés. También empleó restricciones como criterios de decisión. Los dos sitios tratados en este trabajo (GT y AG) proponen reglas bastante robustas. En un organismo como el *P. falciparum*, caracterizado por un genoma rico en AT, era de esperar que las transiciones de contenido AT se hicieran discriminadoras de los sitios de interés. Así ocurrió en algunos casos pero presentando tipos de transición no esperadas. Dado el poco contenido GC del organismo se esperaba que las transiciones de este tipo no jugaran un papel importante, lo que resulto ser cierto. Finalmente, debemos decir que la aproximación por etapas de aprendizaje permite construir un método de predicción en el que los conceptos aprendidos en una etapa pueden ser sometidos a prueba en etapas posteriores, interrelacionándose con otros conceptos para definir descripciones estructurales. Esta aproximación es laboriosa pero confiable pues el conocimiento experto es reevaluado antes de ser incorporado como conocimiento establecido y el resultado es un modelo declarativo.

Trabajo Futuro

La observación del reticulado de hipótesis por el biólogo puede plantearle nuevas estrategias de análisis. Esto último pudiese permitir el replanteamiento de los propios experimentos de aprendizaje, abriendo otras alternativas de exploración y consecuentes metodologías de trabajo. Para colocar al biólogo en ese "loop" de interacción, la interfaz con el inductor debe ser más amigable. También compete una revisión detallada de los criterios empleados para lo que se propone un *text mining* de criterios y restricciones. Es necesario ampliar el número de ejemplos estudiados generándolos a partir de otros cromosomas del organismo en estudio y de los genomas de eucariotas superiores. Es necesario determinar si los resultados son extensibles a otros cromosomas del mismo organismo realizando predicciones basadas en las reglas halladas. Compete además la determinación de modelos para otros organismos. Para desarrollar esto último se pueden tomar estructuras de genes validadas manual o semi-automáticamente disponibles en la comunidad científica (i.e. VEGA, Ensembl) (Searle SM et al 2004; Ashurst JL et al 2005; Ashurst JL and Collins JE 2003; Birney E. et al 2006).

La metodología de trabajo empleada puede complementarse con otras herramientas. Debido a la ausencia de buenos consensos para la zona de ramificación y la zona de poliadenilación, se experimentó el desarrollo de alineamientos generalizados por adeninas, pirimidinas y timinas. La idea es postular consensos que luego sean puestos a prueba por el motor de inducción. Este aspecto, que resulta colateral al presente trabajo, se experimentó de manera muy superficial y merece mayor atención en etapas futuras del trabajo. El reticulado de hipótesis puede ser también una herramienta importante en el análisis de sitios de splicing alternativo. Para ello hay que plantearse el estudio de sitios verdaderos rodeados de sitios aparentemente falsos que pueden resultar muy parecidos a los primeros. Aquellas reglas que expliquen de manera similar a ambos son candidatas a describir las características compartidas por sitios verdaderos que se "agrupan". Este aspecto es de particular interés en la tercera etapa del trabajo.

La tercera etapa propone el desarrollo de experimentos de predicción de intrones. Para ello pueden aprovecharse otras facilidades ya presentes en los conceptos ya aprendidos puesto que la programación desarrollada ha sido conceptualizada considerando facilidades necesarias para las etapas posteriores del trabajo. La cuarta etapa incluirá todo lo aprendido y validado con otros conceptos y restricciones. Deberán incorporarse criterios del tipo: "Todo gen tiene al menos un exón", "La longitud de toda concatenación de zonas codificantes de todo gen es múltiplo de tres" y "todo gen tiene una zona promotora o habilitadora que lo antecede".

Debe realizarse una adecuada comparación de resultados empleando herramientas especializadas en la predicción de sitios de splicing (Majoros et al, 2005; Arita et al 2002; Pertea et al, 2001), procurando simultáneamente mejoras en los procedimientos de evaluación (e.g. x-cross validation)(Ian y Eibe, 2005). Otro aspecto a desarrollar es la exploración de mecanismos para el aprendizaje que incorporen representaciones de la incertidumbre en las reglas. Esto es posible mediante lo que se conoce como programas lógicos estocásticos (Watanabe & Muggleton, 2005) (Muggleton, 2000)(Page, 2000). En nuestro caso esto podría permitir representar reglas y razonar con la incertidumbre inherente o asociable a los conceptos aprendidos en cada etapa.

Agradecimiento
Los autores agradecen el financiamiento otorgado por el FONACIT, bajo el código S1-2000000819.

Referencias

Arita M, Tsuda K and Asai K. (2002); Modeling splicing sites with pairwise correlations, BIOINFORMATICS Vol. 18 Suppl. 2.
Ashurst JL et al (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res. Jan 1;33
Ashurst JL and Collins JE (2003). Gene annotation: prediction and testing. Annu Rev Genomics Hum Genet. 2003;4:69-88.
Badea L. (2003); FUNCTIONAL DISCRIMINATION OF GENE EXPRESSION PATTERNS IN TERMS OF THE GENE ONTOLOGY; Pacific Symposium on Biocomputing 8:565-576.
Birney E. et al. Ensembl (2006). Nucleic Acids Res. 2006 Jan 1;34
Burge, C. B. (1998) Modeling dependencies in pre-mRNA splicing signals, Computational Methods in Molecular Biology, Vol.32, p. 127-163, Elsevier.
Finn P., S. Muggleton, D. Page, and A. Srinivasan (1998). Pharmacophore discovery using the Inductive Logic Programming system Progol. Machine Learning, 30:241-271.
Hogger, Christopher Jhon (1990); Essentials of Logic Programming Clarendon Press. Oxford, 1990
Hsieha S., Chung Y., Lin C., Tang C. (2005); EXONSCAN: EXON Prediction with Signal Detection and Coding Region AligNment in Homologous Sequences; 2005 ACM Symposium on Applied Computing.
Huestis Robert, Fisher Katia (2001); Prediction of many exons and introns in Plasmodium Falciparum chromosome 2, Molecular & Biochemical Parasitololgy, Elsevier, p. 187-199.
Ian H. Witten, Eibe Frank (2005); Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2005.
Keles S., J. van der Laan M., Vulpe C.; Regulatory motif finding by logic regression; BIOINFORMATICS, Vol. 20 no. 16 2004, pages 2799–2811.
King R., Wise P., and Clare A.; Confirmation of data mining based predictions of protein function; BIOINFORMATICS, Vol. 20 no. 7 2004, pages 1110–1118.
Kolchanov N., Pozdnyakov M., Orlov Y., Vishnevsky O., Podkolodny N. (2002); Computer System "Gene Discovery" for Promoter Structure Analysis; In Silico Biology, Issue: Volume 2, Number 3 / 2002, Pages: 257 - 262
Lehninger, Albert L (1999), Principles of biochemistry, Worth Publishers.
Majoros W. H., Pertea M. and Salzberg S. L. (2004); TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, Bioinformatics Volume 20, Number 16 Pp. 2878-2879
Muggleton S.H. (2005); Machine learning for systems biology. In Proceedings of the 15th International Conference on Inductive Logic Programming, LNAI 3625, pages 416-423. Springer-Verlag, 2005.
Muggleton Stephen, Firth Jhon (2003); Cprogol4.4: A tutorial introduction, Computational Bioinformatics Laboratory, Department of Computing, Imperial College, London, United Kingdon. (http://www.doc.ic.ac.uk/~shm/Software/progol4.4/)
Muggleton, S.H. (1995). Inverse entailment and Progol. New Generation Computing, 13:245-286.
Muggleton, Stephen (2000). Learning stochastic logic programs. In Proceedings of the AAAI2000 Workshop on Learning Statistical Model from Relational Data. AAAI, 2000.
Page, David. ILP (2000): Just Do It. J. Lloyd et al. (Eds.): CL 2000, LNAI 1861, p. 25-40.
Padgett, R. A. and Burge, C. B. (2003); Splice Sites. In Encyclopedia of the Human Genome, Nature Press.
Pertea M, Lin X. and Salzberg S. L. (2001); GeneSplicer: a new computational method for splice site prediction,
Nucleic Acids Res. Mar 1;29(5):1185-90
Rawn, J. David (1989), Bioquimica, McGraw-Hill Interamericana, p. 781-820.
Salzberg S.L., Searls D.B., Kasif S., (1998) Grand Challenges in Computacional Biology, Computational Methods in Molecular Biology, Vol.32, p. 3-10, Elsevier.
Searle SM et al (2004). The otter annotation system. Genome Res. 2004 May;14(5):963-70
Siu-wai Leung, Chris Mellish, and Dave Robertson (2001); Basic Gene Grammars and AND-ChartParser for language processing of Escherichia coli promoter AND sequences. Bioinformatics 17: 226-236.
Srinivasan, R.D. King S.H. Muggleton, and M. Sternberg (1997). Carcinogenesis predictions using ILP. In N. Lavrac and S. Dzeroski, editors, Proceedings of the Seventh International Workshop on Inductive Logic Programming, p. 273-287. Springer-Verlag, Berlin LNAI 1297.
Tamaddoni-Nezhad A. , Chaleil R., Kakas A., and Muggleton S.H. (2006). Application of abductive ILP to learning metabolic network inhibition from temporal data. Machine Learning, 2006. DOI: 10.1007/s10994-006-8988-x.
Tom, Mitchell (1997); Machine Learning, ISBN 0-07-042807-7, McGraw-Hill.
Turcotte M, S.H. Muggleton, and M.J.E. Sternberg (1998). Protein fold recognition. In C.D. Page, editor, Proc. Of the 8th International Workshop on Inductive Logic Programming (ILP-98), LNAI 1446, p. 53-64, Berlin. Springer-Verlag.
Turcotte, S.H. Muggleton, and M.J.E. Sternberg (2001). Automated discovery of structural signatures of protein fold and function. Journal.of Molecular Biology, 306:591-605.
Wang Z., Rolish M., Yeo G., Tung V., Mawson M. and Burge, C. (2004); Systematic Identification and Analysis of
Exonic Splicing Silencers; Cell, Vol. 119, 831–845, December 17.
Watanabe H. and Muggleton S.H. (2005). Learning Stochastic Logical Automaton. In Proceedings of the 19th Annual Conferences of JSAI, LNCS 4012, pages 201-211. Springer-Verlag, 2005.
Wen-Hsiung, Li (1997), Molecular Evolution. ISBN 0-87893-463-4 p. 7-34 Sinauer Associates.
Wren J., Johnson D., and Gruenwald L. (2005); Automating Genomic Data Mining via a Sequence-based Matrix Format and Associative Rule Set; BMC Bioinformatics, 6(Suppl 2):S2