List of publications

Publications included in this thesis

1. **Shah PK**, Perez-Iratxeta C, Andrade M and Bork P. Information Extraction from Full Text Scientific Articles: Where are the Keywords? (2003) *BMC Bioinformatics*. **4**(1): 20.

2. Wattarujeekrit T, **Shah PK**, and Collier N. PASBio: Predicate-argument Structures for Event Extraction in Molecular Biology. (2004) *BMC Bioinformatics*. **5**(1): 155.

3. **Shah PK**, Jensen LJ, Boué S, and Bork P. Extracting Transcript Diversity from Scientific Literature. (2005) *PLoS Computational Biology* **1**(1):e10.

4. **Shah PK**, and Bork P. Learning About Transcript Diversity from Scientific Literature with Support Vector Machines. *Bioinformatics (under review)*

Additional publications

5. **Shah PK**, Aloy P, Bork P, and Russell RB. Structural Similarities to Bridge Sequence Space: Finding New Families on the Bridges. *Protein Science* 2005 **14**(5): 1305-14.

6. Perez-Iratxeta C, Astola N, Ciccarelli F, **Shah PK**, Bork P and Andrade MA. A Protocol for the Update of References to Scientific Literature in Biological Databases. *Appl Bioinformatics*. 2003; **2**(3): 189-91.

7. Müller A, Schackert HK, Lange B, Rüschoff J, Füzesi L, Willert J, Burfeind P, **Shah PK**, Becker H, Epplen JT, and Stemmler S. Novel homozygous MSH2 germline mutation in two brothers with colorectal cancer diagnosed at ages 11 and 12 years. *Human Mutation (submitted).*

8. **Shah PK**, Tripathi L, Jensen LJ, Furlong E, Bork P, and Sowdhamini, R. Structure-function Relationships of Eukaryotic Serine Proteases: Specific Analysis of Drosophila Serine Proteases. *Manuscript under preparation*

Data mining from Scientific Literature (summary)

Function annotation in the genomic context is one of the major challenges facing the discipline of Bioinformatics today. Sequences of entire genomes are continuously being deposited in public databases waiting to be analyzed and annotated. Computational methods and data coming out from various types of high-throughput experiments are now being used to assist in functional annotations and knowledge discovery. Published findings mostly analyzing roles of individual genes are used for gene annotations. Similarly, curated sets of facts established in the literature are required in order to check the quality of computational methods and analysis of high-throughput data. Hence, there is a great demand for information extraction tools to extract structured information about gene and gene products from scientific literature automatically and prepare knowledgebases.

Before one sets on to devise tools for information extraction from scientific literature, several questions must be answered. Where does the useful information reside? Is this information structured enough to be extracted? What tools should be utilized for accurate retrieval and extraction of information? Also, how useful mining of information form biomedical texts is for advancing level of present knowledge? Moreover, suitability of tools developed for processing of general English should also be checked for their usability for biomedical texts.

The work presented in this thesis tries to answer questions posed above. Keyword-based analysis of full-text articles from *Nature genetics* was carried out in order to analyze and compare the distribution of information in different sections of papers. Keyword based methods while very useful to explore the overall structure and article contents don't provide exact relationships mentioned in the literature. Biologically important events and relationships can only be extracted using the structured templates based on contents of sentences describing events of interest, which is a non-trivial task. The potential of predicate argument structures for providing semantic templates for accurate information extraction was explored for verbs describing gene expression, molecular interactions and signal transduction. Predicate argument structures (PAS) was defined for important verbs by analyzing sentences from Abstracts as well as full-text articles; they were then compared systematically with PropBank PAS for general English in order to characterize domain specific usage of predicates in biomedical texts.

A database of transcript diversity was generated using a composite procedure that combined retrieval of appropriate sentences from MEDLINE and extracting information using rules based on PAS. Support vector machines proved to be the best sentence categorization/retrieval method when compared to other retrieval methods. LSAT – a database of alternative transcripts was generated after the PAS based information extraction step. Information residing in LSAT was utilized for MeSH term and gene annotations, and studying about the extent of synergy and preference of different transcript diversity generating mechanisms by different organ systems.

Resumen en Español

INTRODUCCIÓN

La anotación de funciones en el contexto genómico es uno de los mayores retos a los que se enfrenta la Bioinformática hoy día. Continuamente, se depositan Secuencias de genomas enteros en bases públicas de datos, esperando a ser analizadas y anotadas. Hoy día se utilizan métodos computacionales y conocimiento procedente del analisis de experimentos de "high-throughput" en la anotación funcional y descubrimiento de nuevo conocimiento.

Para la anotación de genes se utilizan las publicaciones que analizan genes de manera individual. Se necesitan revisiones de hechos publicados en la literatura científica para cubrir las necesidades de conocimiento de científicos individuales, para evaluar la calidad de los métodos computacionales y la cualidad del análisis de datos de "high-throughput". Hay, por lo tanto, una gran demanda de herramientas de procesamiento de lenguaje natural que puedan extraer automáticamente información estructurada sobre genes y sus productos de la literatura científica (Andrade y Bork, 2000; Blaschke *et al*., 2002; Krallinger *et al*., 2005).

Antes de ponerse a diseñar herramientas para la extracción de la información (Information Extraction, IE) presente en los diferentes apartados de un artículo, se debe responder a varias preguntas: ¿Basta utilizar los resúmenes (abstracts) como fuente para la IE, o se debe considerar todo el texto? ¿Dónde reside la información útil dentro de todo el texto de un artículo? ¿Es esta información diferente en diferentes apartados, y esta además suficientemente estructurada para ser extraída?. También: ¿Cuán útil puede ser para incrementar el nivel de conocimiento actual la extracción automática de información de textos biomédicos?. Más aun, se debería comprobar si las herramientas generales de procesamiento del inglés común también pueden ser utilizadas para textos biomédicos. En el trabajo que se presenta a continuacion se intenta dar respuesta a estas preguntas analizando el resúmem (Abstract) y el texto completo de textos biomédicos empleando varias herramientas derivadas de la tecnología de procesamiento del lenguaje natural ("Natural language processing technology").

RESULTADOS

1-Análisis de artículos completos mediante palabras clave

Los resultados de este apartado estas descritos en el siguiente artículo (Shah et al., 2003).

Metodos: Definiendo las palabras clave

El objetivo del trabajo es comparar la información presente en distintos apartados de un artículo, especialmente la diferencia entre el Resumen (Abstract) y el resto del texto. Para ello, se emplearon un total de 104 artículos de la revista *Nature Genetics*, que contienen una estructura regular, a saber: Resumen, Introducción, Métodos, Resultados y Discusión (A, por Abstract, I, M, R y D, respectivamente). Para

simplificar, el trabajo se centra en la extracción de palabras relevantes (palabras clave, o "keywords"), que son palabras que presentan una visión lógica de un documento dado. Para derivar las palabras clave de un apartado de un artículo, se analizaron computacionalmente las asociaciones entre las palabras de dicha sección. Las oraciones se tomaron como la unidad de texto en la que buscar las asociaciones. Se asumió que dos palabras estaban asociadas en el contexto de un apartado si aparecían conjuntamente de manera repetida en oraciones dentro del mismo. Se diseñó un esquema de valoración que daba una puntuación [K] mayor a palabras con muchas relaciones con otras palabras. En este análisis sólo se consideraron palabras definidas como *nombre*.

Resultados

Selección de palabras clave por apartado

El número de palabras seleccionadas que superan un umbral de K varía en diferentes apartados. Encontramos un pequeño número de palabras cuyo valor K era muy superior al resto; esto significa que la organización de las palabras posibilita extraer palabras clave para los cinco apartados considerados. El número de palabras seleccionadas fue muy similar para todos los apartados, para valores muy altos de K (superiores a 0,8). Para un umbral de K >= 0,5, el número resultante de palabras clave fue bastante similar para la Introducción y los Métodos (alrededor de 15 cada una), teniendo cada uno de los otros tres apartados unas nueve palabras clave. Sin embargo, si se tiene en cuenta el tamaño de los apartados, es obvio que la frecuencia más alta de palabras clave por nombre (seleccionadas con K >= 0,5) se alcanza en el Resumen (0,18), seguida por la Introducción (0,08), y quedando después Métodos, Resultados, y Discusión. Esto justifica las estrategias de extracción de datos (o "data mining") que se limitan a analizar los resúmenes para minimizar el trabajo computacional; y sin embargo, nuestro resultado indica que no todas las palabras clave están en el Resumen, y que por tanto podría valer la pena analizar el resto del texto.

Heterogeneidad de información entre los distintos apartados

Con el fin de estudiar la hetogeneidad de la información presente en los distintos apartados, se examinaron aquellas palabras clave en común entre apartados. Los resultados indican que no muchas palabras clave están presentes en todos los apartados, y que aquéllas que lo están no son muy relevantes. Incluso para un umbral bajo de K (K >= 0,3), había una media de sólo una palabra clave general por artículo. Éstas suelen ser vocablos no informativos como "gen" o "proteína". Esto indica que la información no está homogéneamente distribuida entre los apartados de un artículo; es decir, distintos apartados contienen distintos tipos de información.

Para cuantificar las diferencias y similitudes de contenido a lo largo del artículo, se comparó el número de palabras clave compartidas entre apartados diferentes. Los valores indican que la sección Métodos es la más diferente de todas: El contenido de Métodos suele centrarse en las técnicas y protocolos utilizados, y no tanto en el fenómeno biológico tratado en el artículo. Esto de por sí ya explica por qué las palabras clave presentes en esta sección ("proteína" o "gen", por ejemplo) son escasas y carecen de interés.

Respecto a las similitudes entre apartados, los niveles de similitud entre A, I y D son semejantes, y R es el más cercano a M, como se muestra en un dendograma basado en una matriz de distancias (Figura 1). Si algún apartado tiene que tratar sobre los métodos utilizados, aparte del propio Métodos, es precisamente el de Resultados, porque ahí los procedimientos utilizados son relevantes. La Discusión se centra de nuevo en los resultados biológicos (haciendo énfasis en su relación con el conocimiento previo, expuesto en la Introducción) sin entrar en detalles sobre las técnicas ya explicadas en Métodos y justificadas en Resultados. Esto indica que cada apartado contiene ciertas palabras clave que son únicas del apartado. A continuación intentamos caracterizar las diferencias de contenido entre apartados.

Análisis cualitativo de temas por apartado

Un conjunto de palabras (no necesariamente seleccionadas como palabras clave) presentes en el cuerpo de 104 artículos se clasificó en siete grupos para hacer un análisis más profundo del tipo de información residente en cada uno de los apartados. Para hacerlo del modo menos ambiguo posible, se utilizaron las palabras (nombres) que encajaban en las descripciones MeSH (Medical Subject Headings) para esa misma palabra y que pertenecían únicamente a una de las categorías principales de MeSH (que son: Anatomía, Organismos, Enfermedades, Compuestos y Drogas, Técnicas y Equipamiento, y Ciencias Biológicas. Se definió una categoría adicional X en este trabajo: Unidades, Dimensiones y Partes). Se estudió el número medio de aparición y densidad de las palabras de cada uno de los siete grupos;

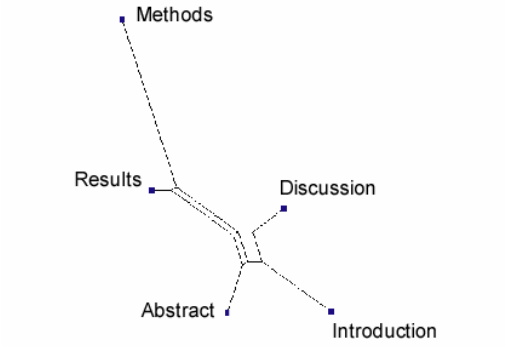

Figura 1 – Comparación entre apartados según el contenido en palabras clave: Se muestra gráficamente la similitud entre apartados de acuerdo con el contenido en palabras clave.

los resultados indican que los apartados de un artículo son una buena fuente de palabras clave. El Resumen parece ser la mejor fuente para la mayoría de los temas, con respecto a la frecuencia de palabras clave, excepto para aquellos temas típicos de la sección de Métodos (Técnicas y Equipamiento; Compuestos y Drogas). Introducción, Resultados y Discusión contienen una gran cantidad de información relacionada con enfermedades, y Métodos tiene muchos términos relacionados con técnicas.

2-IE de propósito general, y extracción de eventos

Este apartado se resume en el siguiente artículo (Wattarujeekrit *et al*., 2004).

Los métodos de IE basados en palabras clave proporcionan información sobre el contenido del texto estudiado. Sin embargo, no se pueden utilizar para crear tablas estructuradas en bases de datos; para ello se requieren herramientas de IE que puedan encontrar los eventos o relaciones exactos que se describen en el texto. El objetivo de la IE es proporcionar unidades de conocimiento estructurado a partir de texto libre no estructurado, combinando aproximaciones desde áreas tales como el procesamiento de lenguaje natural y el aprendizaje de máquinas. La extracción de eventos funciona mediante el uso de registros y campos predefinidos, de acuerdo con un contexto particular. Sin embargo, el rendimiento de los métodos de IE que utilizan reglas basadas en la sintaxis de las oraciones disminuye por el hecho de que una frase se puede escribir de muchos modos diferentes y gramaticalmente correctos. El problema de los patrones sintácticos se encuentra en toda suerte de textos, incluidos los científicos (Figura 2).

La necesidad de relaciones semánticas en la extracción de eventos moleculares

A continuación se ilustra con un ejemplo la necesidad de relaciones semánticas en la extracción de eventos moleculares.En las oraciones describiendo el evento expresión (Figura 2) los campos de información son: A – entidad expresada, B – propiedad física de la entidad expresada, y C – localización, referida al orgánulo, célula o tejido. En la oración 1 (donde A = la enzima, B = dos isoformas de mRNA de 2,4 y 4,0 kb, y C = encéfalo), la información necesaria para describir el evento con respecto al campo B se distingue utilizando un sintagma preposicional; en cambio, en la oración 2 se utiliza una aposición (donde A = dos mRNAs para *il8ra* igualmente abundantes, B = de 2,0 y 2,4 kb de longitud, C = neutrófilos), sin que ello tenga trascendencia en la descripción del evento en que participa. La oración 3 (donde A = RNA y proteína para los cuatro TCR transgénicos, y C = células T, sin mencionar B) ilustra otro problema, esta vez concerniente a “células T”, porque desde una perspectiva biológica “células T” valdría igualmente como fuente o localización, no sólo como un agente desde el punto de vista lingüístico.

(1) El análisis por Northern blot con mRNA de ocho tejidos humanos diferentes mostró que [la enzima $_{A}$] se expresaba exclusivamente en [el encéfalo $_{C}$], con [dos isoformas de mRNA de 2,4 y 4,0 kb $_{B}$].
(2) [Dos mRNAs para *il8ra* igualmente abundantes $_{A}$], [de 2,0 y 2,4 kb de longitud $_{B}$], se expresan [en neutrófilos $_{C}$], y surgen del uso de dos señales alternativas de poliadenilación.
(3) Esta “exclusión alélica funcional” se debe aparentemente al control del proceso de ensamblaje del TCR, porque estas [células T $_{C}$] expresan [RNA y proteína para los cuatro TCR transgénicos $_{A}$].

Figura 2 – Ejemplo de diferentes formas de *Expresión*: La variación superficial de expresiones lingüísticas para el evento *expresión* es clara en las oraciones (1) a (3). La oración 3 enfatiza el hecho de que se requiere conocimiento especializado para comprender el significado de la oración (ver el texto).

Estos ejemplos muestran que el uso de expresiones sintácticas regulares en textos superficiales no sería adecuada para una buena IE, dada la complejidad en estructuras superficiales. Por tanto, un método fiable de IE debiera resolver los problemas de dependencia del contexto y multipatrón sintáctico. Tratar con estos problemas requiere explotar el conocimiento estructural y semántico en la profundidad de las oraciones del texto bajo análisis. Estos requerimientos pueden satisfacerse si se consigue agrupar varias estructuras superficiales en una misma estructura predicativa (Predicate-Argument Structure, PAS), representando la información con argumentos, los roles semánticos que juegan las distintas entidades junto a un verbo que comunica un evento concreto.

Métodos

Estructuras predicativas (Predicate-Argument Structures, PAS)

Con la intención de proporcionar a la "comunidad bio-IE" una fuente fiable de PAS, se preparó una base de datos (PASBio) de predicados frecuentemente utilizados en el área de regulación de la expresión génica, interacciones moleculares y transducción de señales (Wattarujeekrit *et al*., 2004). La metodología de PASBio se tomó de PropBank (Kingsbury and Palmer, 2002; Kingsbury *et al*., 2002), la base de datos de PAS para el Inglés general, con las adaptaciones apropiadas. Para definir una PAS para cada verbo, se hizo una exploración del uso del verbo y el acompañamiento de distintos argumentos a partir de una muestra de oraciones procedentes de resúmenes (abstracts) y de artículos enteros. Un verbo podía tener varios significados según su uso (por ejemplo, "express" para "hablar" o para "envío rápido"). En PASBio se dividieron estos significados con el objetivo de obtener sentidos semánticos unitarios; para ello se utilizó el diccionario WordNet (Miller, 1990). Cada registro PAS en PASBio contiene un conjunto de argumentos fundamentales, y argumentos auxiliares. Un argumento se considera fundamental si es importante para completar el significado del evento descrito en la oración, a los argumentos fundamentales se les asignan unas etiquetas *ArgX* (donde X es un número cardinal, comenzando en 0 e incrementándose con cada argumento adicional) y *ArgR*, además de las etiquetas mnemónicas que tratan sus roles biológicos.

Resultados

Algunas conclusiones del análisis son las que siguen: a un argumento se le debería asignar la etiqueta ArgX si es un argumento fundamental (desde el punto de vista de la IE) y su rol se justifica durante el evento dictado por el predicado. Al argumento que tiene un rol después del evento se le tiene que asignar la etiqueta ArgR (de "resultado"). Los predicados encontrados en textos biomédicos son normalmente específicos del campo de estudio; es más, tienen conjuntos de argumentos distintos de los que se requieren para los predicados del Inglés general. Los argumentos de un predicado no sólo completan la descripción de un evento, sino que además pueden modificarlo completamente con su presencia. Argumentos con roles como agente, instrumento o localización son comunes entre PAS de textos biomédicos e Inglés general. Los roles biológicos de un argumento pueden diferir de sus roles lingüísticos.

La base de datos PASBio, que contiene las PAS de predicados de textos biomédicos, está disponible en http://research.nii.ac.jp/~collier/projects/PASBio. Aunque PASBio se diseñó para ser utilizada como un diccionario semántico específico de textos biomédicos para una IE precisa, se puede utilizar en cualquier aplicación que requiera obtener la forma lógica de una oración dada. Tales aplicaciones incluyen aprendizaje de máquinas sobre etiquetado semántico de roles, traducción automática, y confección automática de resúmenes.

3-IE sobre la diversidad de tránscritos

El trabajo que se describe en este apartado se resume en los siguientes artículos (Shah *et al*., 2005 and Shah y Bork, en revisión).

La generación de diversidad de tránscritos por "splicing" alternativo (Alternative Splicing, AS) y mecanismos asociados contribuyen enormemente a la complejidad funcional y a la evolución de los sistemas biológicos (Boue et al., 2003). Los numerosos ejemplos de los mecanismos y sus implicaciones funcionales se encuentran dispersos en la literatura científica. Por tanto, es crucial tener una herramienta que pueda extraer los hechos relevantes automáticamente y reunirlos en una base de conocimiento, lo que puede ayudar en la interpretación de datos de los métodos de "high-throughput" y asentar una base más firme para el desarrollo de futuras herramientas computacionales.

Métodos

Estrategia general para la generación de la base de datos de diversidad de transcritos a partir de la bibliografía

Se diseñó un procedimiento de dos pasos para extraer la información dispersa en MEDLINE sobre diversidad de transcritos y su expresión espacio-temporal. En el primer paso se identificaron las oraciones con información sobre diversidad de transcritos en los resúmenes de los artículos. Para ello, y para sortear el problema de los patrones sintácticos, un conjunto de clasificadores fue entrenado para identificar dichas oraciones sobre diversidad de transcritos; los clasificadores estaban basados en distintos algoritmos de categorización de texto, y aprendieron con un método inductivo. El mejor clasificador fue entonces utilizado para procesar la base de datos MEDLINE entera, identificando unos 14000 resúmenes con oraciones describiendo diversidad de tránscritos. En el segundo paso se dividieron las oraciones en sus constituyentes, y estos se distribuyeron en ocho categorías semánticas diferentes (etiquetas de argumento. Tabla 1).

La información sobre genoma, transcrito y secuencias de proteínas se asoció a los identificadores de PubMed correspondientes utilizando las referencias bibliográficas en bases de datos como Swiss-Prot (Bairoch y Apweiler, 2000), Refseq (Pruitt and Maglott, 2001), GenBank (Benson *et al*., 2004), y Ensembl (Birney *et al*., 2004) cuando fue posible. Por tanto, cada registro en la base de datos LSAT (Literature Support for Alternative Transcripts) contiene el título del artículo, el resumen, categorías semánticas extraídas de las oraciones, y referencias a otras bases de datos. Esta base de datos contiene, en resumen, 3063, 769, 105 y 207 ejemplos no redundantes de "splicing" alternativo, uso diferencial de promotor, y

poliadenilación alternativa extraídos de la bibliografía y asociados con genes, tejidos y especies. Además, los casos de uso alternativo de promotor con nombres de genes y tejidos extraídos en este trabajo son la mayor colección de este evento disponible hasta la fecha. Esta colección sería útil en el análisis de regiones promotoras. LSAT está disponible en http://www.bork.embl-heidelberg.de/LSAT/.

Rendimiento en clasificación de oraciones y extracción de información

Se comparó el rendimiento de la clasificación de oraciones que describen la generación de diversidad de tránscritos (con distintas fracciones del conjunto de entrenamiento) con los siguientes métodos de clasificación: 1) "naive Bayes", 2) entropía máxima, 3) "expectation maximization", 4) "k-nearest neighbor", 5) variantes del "term-frequency inverse document frecuency", y 6) "Support vector machines" (SVM). Además se generaron, a partir de los conjuntos de entrenamiento, cuatro grupos distintos de rasgos de aprendizaje, con diferentes niveles de riqueza de rasgos.

El SVM mostró un rendimiento superior a todos los demás en la clasificación de oraciones cuando se le entrenó con una "bag of words" como conjunto de entrenamiento. Es más, un SVM con un núcleo de función de base radial (Radial Basis Function, RBF) rindió mucho más que SVMs con núcleo linear o sigmoide. El clasificador final fue entrenado con valores gamma de 1,5 y C de 10, y "bag of words and phrases" como conjunto de rasgos, tras una elaborada optimización de parámetros. Este clasificador alcanzó una precisión del 66 % y un "recall" de 74.33 al aplicarlo sobre el MEDLINE entero. La precisión y el "recall" para identificar varias categorías semánticas se muestra en la tabla 1.

Semantic Category	Presence (%)	Recall (%)	Precision (%)	Total Instances
Event mechanism	79	92	96	13103
Gene names	71	82	88	15905
Tissues	22	87	96	5028
Species	21	97	99	4093
Number of isoforms	20	77	100	2965
Diff. In structure/function	12	63	86	1620
Experimental methods	11	57	82	1071
Specificity	5	100	85	1589

Table 1 – Rendimiento a la hora de extraer categorías semánticas

Resultados

Cuantificación de los distintos mecanismos que llevan a diversidad de transcritos

Mientras se analizaban las oraciones etiquetadas con varias categorías, se encontró que el uso diferencial de promotor se daba junto con "splicing" alternativo (AS) en un 12 % de los resúmenes. El 19 % de los resúmenes que trataban un uso alternativo de primer exón también mencionaban el uso de

diferentes promotores. Un 17 % de los resúmenes que describían una poliadenilación alternativa también mencionaban un AS. El alcance descrito aquí de esta sinergia entre mecanismos es probablemente una subestimación del alcance real, pues el clasificador detecta menos casos de uso diferencial de promotor o de poliadenilación alternativa que casos de AS (y en la bibliografía sucede lo mismo, describiéndose mucho más el AS que los otros dos fenómenos).

El peso de cada uno de los mecanismos de generación de diversidad de transcritos podría variar según el sistema anatómico y la etapa del desarrollo (Figura 3a; panel superior). Para estudiar dicha posibilidad se estudió en qué organos tenían lugar todos los distintos eventos extraídos de la bibliografía (limitándose a vertebrados), teniendo en cuenta genes y tejidos; se utilizaron para esto los términos anatómicos MeSH. La figura 3 muestra que los cuatro mecanismos de generación de diversidad de transcritos se utilizan igualmente en la mayoría de sistemas. Sin embargo, había una representación significativamente superior (Figura 3a, panel inferior) de AS en el sistema nervioso, sugiriendo que existe una preferencia por este mecanismo en este sistema. Del mismo modo, había una gran frecuencia de uso diferencial de promotor en tejidos conectivos, y en menor grado en el aparato digestivo y los genitales.

Diferencias específicas de tejido en el alcance del "splicing" alternativo

Disponiendo de una gran cantidad de eventos de AS de alta calidad, las diferencias específicas de tejido para el AS debieran ser visibles. Se ha demostrado un papel importante del AS en causar especializaciones funcionales en tejidos y etapas del desarrollo (Grabowski and Black, 2001; Yeo et al., 2004). Se analizaron manualmente los registros en LSAT conteniendo el campo "especificidad". Tras una revisión de la información que faltaba sobre identificador génico y tejidos, encontramos 959 eventos describiendo un "splicing" específico de tejido. Los resultados incluían 400 eventos no redundantes para 183 genes humanos. 190 genes más de varias especies fueron también asociados a identificadores de Swiss-Prot durante la revisión manual.

Para estudiar el alcance del AS específico de tejido, agrupamos como antes los órganos y tejidos en los sistemas respectivos, y representamos (Figura 3b, panel izquierdo) el alcance observado de AS mediante intensidad de colores. El sistema nervioso (L), los genitales (H), el sistema inmune (I), el aparato digestivo (D) y el músculo esquelético (K) mostraron una gran especificidad de "splicing", tanto dentro de un mismo sistema como entre sistemas. Hay también casos de transcritos obtenidos por AS exclusivos de un sistema, siendo el sistema nervioso el que mostraba la mayor cantidad de estos transcritos únicos. Estos patrones de expresión específicos de tejido extraídos de la bibliografía solapan en gran medida con los 667 eventos de AS específicos de tejido que se dedujeron de los datos de ESTs (Xu et al., 2002) de 454 genes humanos en 46 tejidos (Figura 3b, panel derecho).

El conocimiento extraído de la bibliografía confirma, como también lo hicieron antes ciertos trabajos experimentales (Mirnics and Pevsner, 2004), los estudios basados en ESTs (Xu *et al*., 2002; Yeo *et al*., 2004), que también muestran el uso del AS como mecanismo prevalente en la generación de diversidad de transcritos en el sistema nervioso. Estudios basados en ESTs (Yeo *et al*., 2004) también sugirieron que genes del hígado (aparato digestivo) y los testículos (genitales) muestran distintos patrones de "splicing"

con exones alternativos. Nuestros resultados indican que estos transcritos podrían mostrar estos patrones diferentes de "splicing" en combinación con distintos promotores. Esta conclusión parece plausible si se tiene en cuenta que el AS de exones terminales se ve influenciado por promotores alternativos en al menos un 19 % de los casos (resultados arriba; (Zavolan *et al*., 2003)), y se debería seguir explorando.

También se utilizó el conocimiento en LSAT para asignar el término MeSH "alternative splicing" a los 1536 resúmenes en MEDLINE que debieran tenerlo pero carecían de él; también se proporcionaron anotaciones con respecto a transcritos alternativos para 1860 genes en Swiss-Prot y Refseq y transcritos generados *de novo*.

CONCLUSIONES

1.- Hay una necesidad clara de realizar extracción de información de datos biológicos sobre el texto completo de artículos científicos. La distribución de información en todo el texto de los artículos científicos es heterogénea, y hay una cierta correspondencia entre las secciones del artículo y los distintos tipos y la densidad de datos relevantes.

2.- Los resúmenes (abstracts) de los artículos de ciencias biomédicas son el mejor repositorio desde el punto de vista de densidad de palabras clave, y están disponibles en MEDLINE, justificando los métodos de extracción de información que utilizan sólo los resúmenes. Sin embargo, hay mucha más información relevante en el resto del artículo, especialmente en las secciones de Introducción y Discusión. Es más, la información está suficientemente estructurada como para obtener un gran número de palabras clave.

3.- El análisis de oraciones en resúmenes y texto completo de artículos biomédicos muestra una clara necesidad de utilizar conocimiento semántico para una extracción de información precisa. Los registros PAS (Predicate-Argument Structure) formalizan la definición de modelos (templates) de extracción proporcionando estructuras argumento, las cuales complementan un predicado que describe un evento. Además, el conocimiento semántico residente en los registros PAS ayudará a los procesos de extracción basados en modelos a resolver el problema de múltiples patrones sintácticos.

4.- El uso de predicado en los textos biomédicos es específico de dominio o campo de estudio, y por tanto se necesita una aplicación PAS específica de dominio para una IE precisa. La utilización de PAS también permitirá la creación de un sistema de IE de propósito general para los textos biomédicos. La base de datos PASBio generada como parte de este trabajo es prometedora para estas funciones (disponible en http://research.nii.ac.jp/~collier/projects/PASBio/).

5.- La generación y regulación de transcritos alternativos es un evento importante para la diversidad funcional y la evolución de los eucariotas. Una base de datos de transcritos alternativos (LSAT) fue generada semiautomáticamente utilizando un procedimiento compuesto que contenía identificación de oraciones y pasos de extracción de información. LSAT está disponible en http://www.bork.embl-heidelberg.de/LSAT/

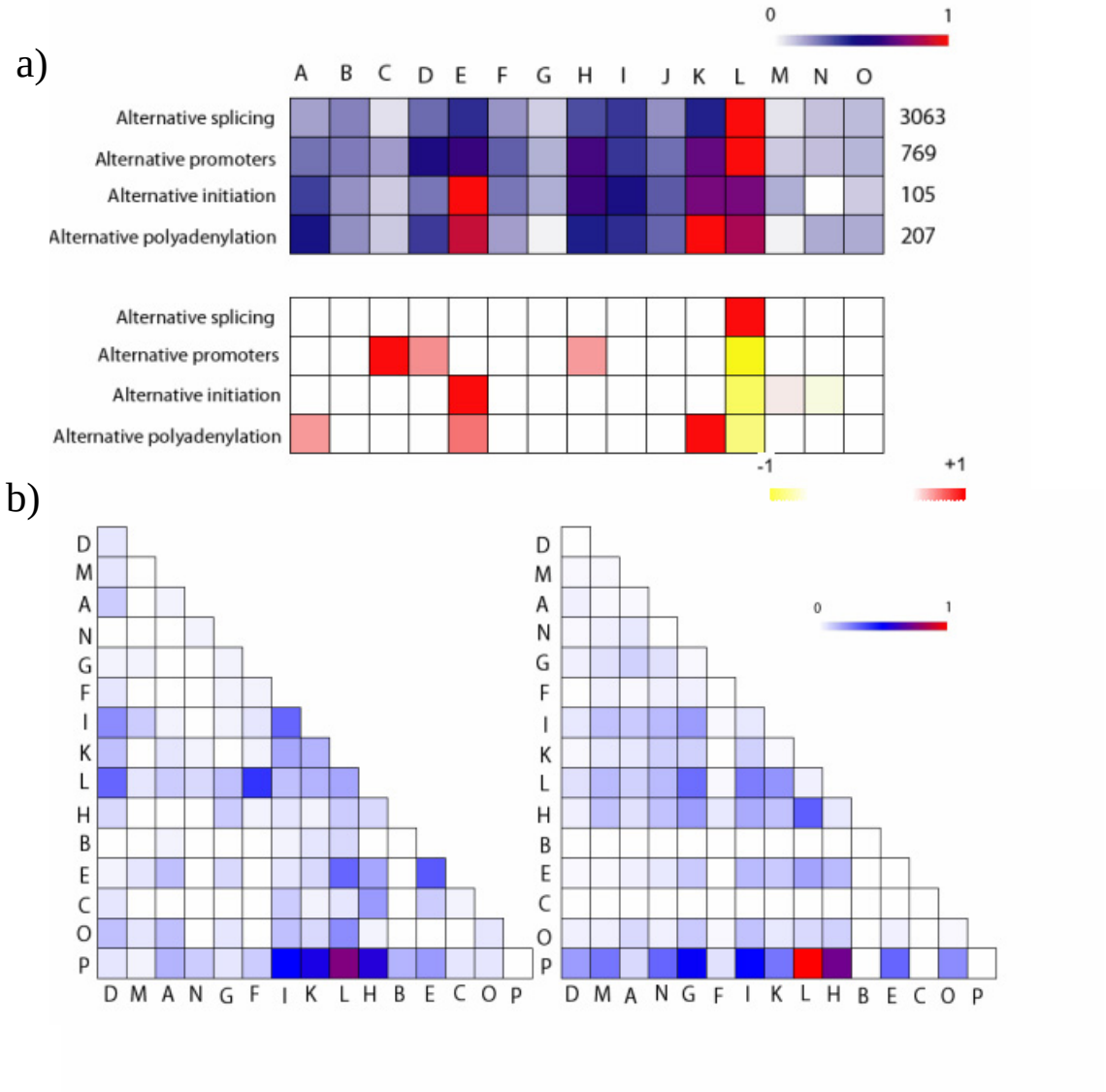

Figura 3: (a) Alcance del uso de varios mecanismos generadores de diversidad. (b) El alcance de "splicing" específico de tejido observado. Se situaron los tejidos en sistemas corporales según la clasificación MeSH. Están señalados con las letras: A: Sistema (sis) cardio-vascular; B: Células; C: Tejidos conectivos; D: Aparato digestivo; E: Estructuras fetales o embrionarias; F: Sis endocrino; G: Glándulas exocrinas; H: Genitales; I: Sis inmune; J: Sis integumentario; K: Sis muscular esquelético; L: Sis nervioso; M: Aparato respiratorio; N: Regiones sensoriales; O: Sistema urinario.

.

6.- Un clasificador basado en “Support vector machines” (SVM) seguido de entropía máxima superó a los otros métodos de clasificación de oraciones. SVM con un núcleo de base radial generalizaba bien; son los mejores clasificadores de los datos de texto. Un aprendizaje automático de clasificación de oraciones también permitió evitar el problema de múltiples patrones sintácticos. Ambos, la clasificación de oraciones y los pasos de extracción de información, alcanzaron una buena medida F en el proceso de “benchmarking”.

7.- LSAT tiene gran cantidad de conocimiento, que fue utilizado para la asignación automática de términos MeSH y anotaciones de función, tanto a genes en bases de datos de secuencias como a transcritos alternativos generados *de novo*.

8.- La búsqueda de datos (“data mining”) de LSAT también permitió poner hipótesis a prueba. Los resultados de prueba de hipótesis y la comparación con datos de ESTs sugieren que el “splicing” alternativo podría ser el mecanismo preferente de generación de transcritos alternativos en el sistema nervioso. Por tanto, el “text mining” no sólo ayuda a analizar datos de otras fuentes, sino que además es en sí mismo una fuente independiente.

Abbreviations

A: Abstract

AP: Alternative polyadenylation

AS: Alternative splicing

D: Discussion

DP: Differential promoters

EM: Expectation maximization

FDG: Functional dependency grammar

FN: False negative

FP: False positive

I: Introduction

IE: Information extraction

LSAT: Literature support for alternative transcripts

M: Methods

ML: Machine learning

MUC: Message understanding conference

NLM: National library of medicine

NLP: Natural language processing

PAS: Predicate argument structure

R: Results

RBF: Radial basis function

SVM: Support vector machines

TD: Transcript diversity

TN: True negative

TP: True positive

List of Figures

List of Tables