Universidad autónoma de Madrid

Facultad de Ciencias

Departmento de Biología Molecular

Data mining from Scientific Literature

TESIS DOCTORAL

Parantu K Shah

European Molecular Biology Laboratory, Heidelberg, Germany

Max-Delbrück Centrum für Molekulare Medizin, Berlin, Germany

Universidad Autónoma de Madrid, Madrid, Spain

Universidad Autónoma de Madrid

Facutad de Ciencias

Departmento de Biología Molecular

Data mining from Scientific Literature

Memoria presentada para optar al grado de Doctor en Biología Molecular por:

Parantu K Shah

Director: Dr. Peer Bork **Tutor: Dr. Alfonso Valencia Herrera**

Acknowledgements

I spent four productive and fun-filled years at EMBL growing academically and personally. I would like to thank everybody who has been part of this experience.

- First and foremost, Dr. Peer Bork for giving me the opportunity to work under his guidance. Peer provided scientific advice, ideas and resources so that the work presented here could be completed. But above all, he provided me with ample freedom to work in still developing field of natural language processing (NLP) of biomedical texts and encouraging me to *push the limits* of the current knowledge.

- Dr. Miguel Andrade and Dr. Carolina Perez-Iratxeta in helping me getting started in NLP in biology.

.

- Dr. Nigel Collier at the NII in Japan for teaching me basics of NLP and for providing me with an opportunity to work under his guidance and interact with his group composed of theoretical and computational linguists.

- Dr. Rob Russell for his useful comments as a member of my advisory committee and giving me opportunity to work under his guidance in the area of Structural Bioinformatics. Other members of my advisory committee, Dr. Christos Ouzounis and Prof. Alfonso Valencia.

- Dr. Lars Juhl Jensen for enlightening me towards the power of statistics and data analysis, and helping me deal with scientific writing and reviewers' reports, the other side of scientific world.

- Dr. Francesca Ciccarelli and Dr. Tobias Doerks for being wonderful colleagues to share the office with. Dr. Monica Campillos for helping me out with the submission procedure. Also, present and past members of Bork group and members of EMBL Biocomputing program for making EMBL a place full of fun and lots of science.

- Andrés Gaytan and Dr. Monica Campillos for helping me with the Spanish version of the summary of thesis work.

- My friends Carolene Lemerle, Maria-Vittoria Verga-falzacappa, Magdalena Kraus, Gabriela Zuliani, Sumati Mattoo, Sonal Patel, Francesca Diella, Peter Winn, Marina Chekulaeva, Barbara Diventura, Fabiana Perocchi, Sandra Esteras, Lodovica Borghese, Rune Linding, Tuangthong

Wattarujeekrit, Tony Mullen, Luree Schneider, Anirban Bhaduri, Elham Andaroodi, Jumpot Phuritatkul, Pedro Beltrao, and others for touching my life in a positive way and enriching it immensely (names are in no particular order, some may be invariably missing).

- My parents and other members of my family for their unconditional love and support. For teaching me to fight for all things I deserve, believe in myself and keep a balanced head.

- My brother for teaching me to work hard and concentrate my energies to the positive direction in life. This thesis is one of the examples of things that can be achieved with a positive outlook.

Index

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;

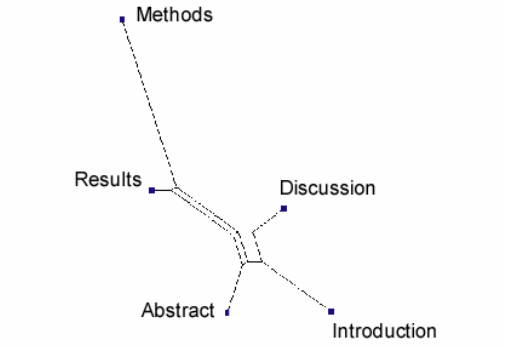

Figura 1 – Comparación entre apartados según el contenido en palabras clave: Se muestra gráficamente la similitud entre apartados de acuerdo con el contenido en palabras clave.

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/

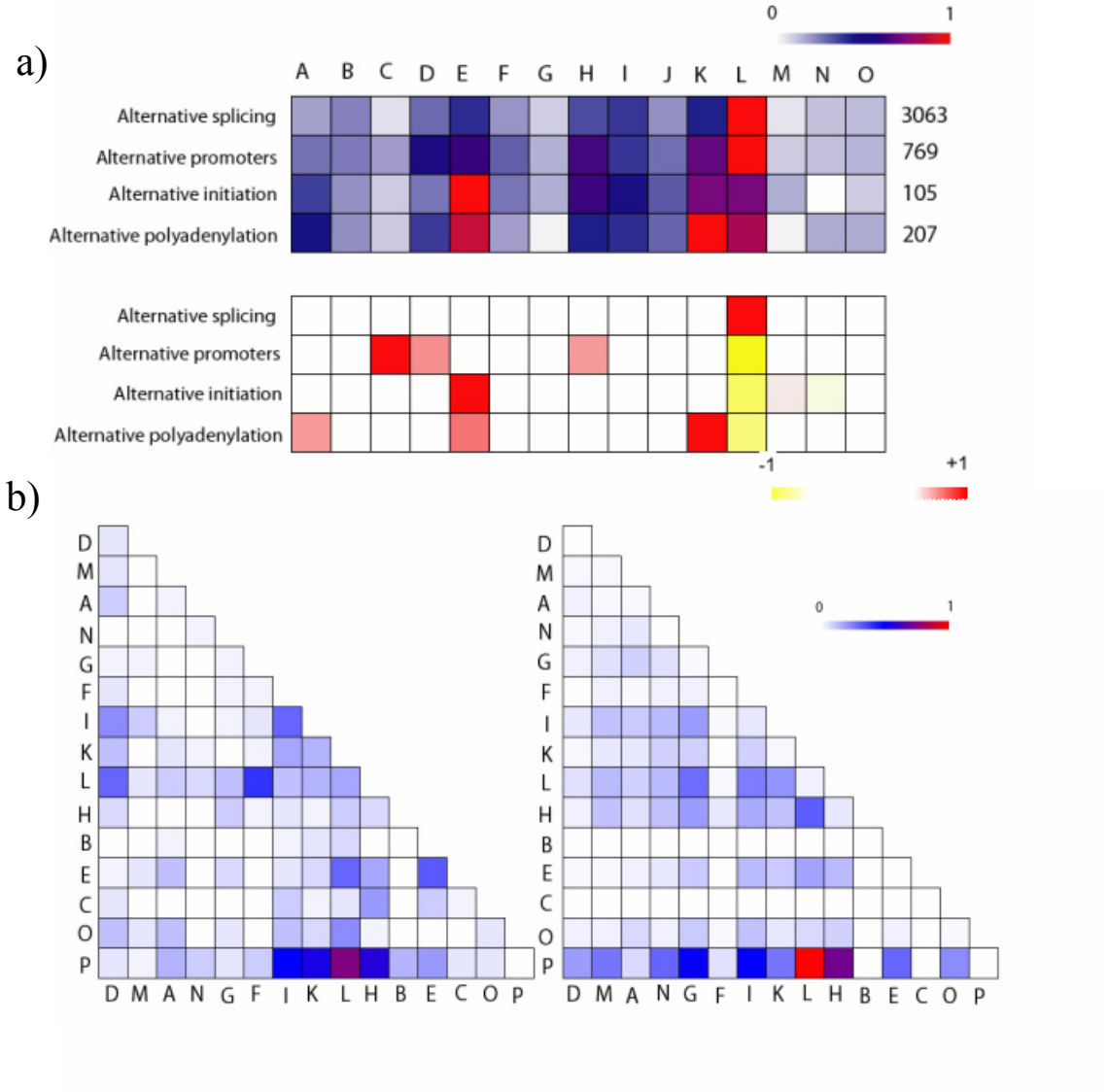

Figura 3: (a) Alcance del uso de varios mecanismos generadores de diversidad. (b) El alcance de "splicing" específico de tejido observado. Se situaron los tejidos en sistemas corporales según la clasificación MeSH. Están señalados con las letras: A: Sistema (sis) cardio-vascular; B: Células; C: Tejidos 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

I. - Introduction

1.1. – Prologue-Introduction

Published literature is the largest and most valuable repository of biological information. A vast majority of these articles describe studies carried out on individual genes. This information is mostly recorded in the form of free-text articles readable by humans and accessible by machines mostly through shallow keyword-based search engines. On the other hand, the availability of complete genome sequences of many model organisms, and data from high-throughput experimental methods are making it possible to design experiments at the whole genome level, ask complex questions and increase the pace of biological discovery.

It is now widely accepted that data generated using high-throughput methods have very high error rates and the knowledge derived from the analysis of the data needs to be assessed against the known facts from the literature (von Mering et al., 2002). Many systems biology approaches based on data integration like STRING have started incorporating the published knowledge as an independent evidence type for building probabilistic gene networks (von Mering et al., 2003) Moreover, many existing databases like SWISSPROT (Bairoch and Apweiler, 2000), and OMIM (Hamosh et al., 2005) associate literature with molecular entities. These databases contain a higher level of relationships, are therefore more informative and can be mined for further knowledge discovery. However, manual curation of these databases is limiting their growth and reducing the accuracy of the information provided. Thus, there has been a surge of interest in using biomedical literature to accomplish different tasks, varying form modest task like finding reported gene location on chromosome to more ambitious attempts to construct putative gene networks (Hoffmann et al., 2005).

Regardless of the explicit goal, there are several major hurdles to overcome when using the biomedical literature for finding information. The most obvious is the shear number of available articles, which is continuously growing. For instance, the most widely used biomedical literature database, NCBI's PubMed, contains over 12 million abstracts but this database by no means covers all the publications in all areas related to biomedicine (Figure 1.11). Another major problem arises when searching for the literature relevant to specific entities such as a gene, a protein, or a disease. Since both the English language and the biomedicine jargon suffer from several levels of ambiguity, the method may miss relevant papers, as well as retrieve irrelevant ones. Yet another issue is the inherent difference between the text that is typically searched by current text handling tools and the scientific literature. Much of the work on text mining aims at, and is tested on, articles such as news reports, typically written by professional writers with aim to convey a story to masses. In contrast, scientific documents are written by scientists whose first language may often not be English, whose primary qualification is research rather than report writing, and whose target audience is a relatively small group of fellow scientists, all familiar with the same domain-specific jargon. Scientific articles thus often use unexplained but widely understood concepts, non-standard terms

and grammatical structures, and include material and background information that may not directly pertain to or may even contradict the paper's main aim point (Netzel et al., 2003). All these factors add a level of complexity to the scientific literature, making it harder to mine with standard tools.

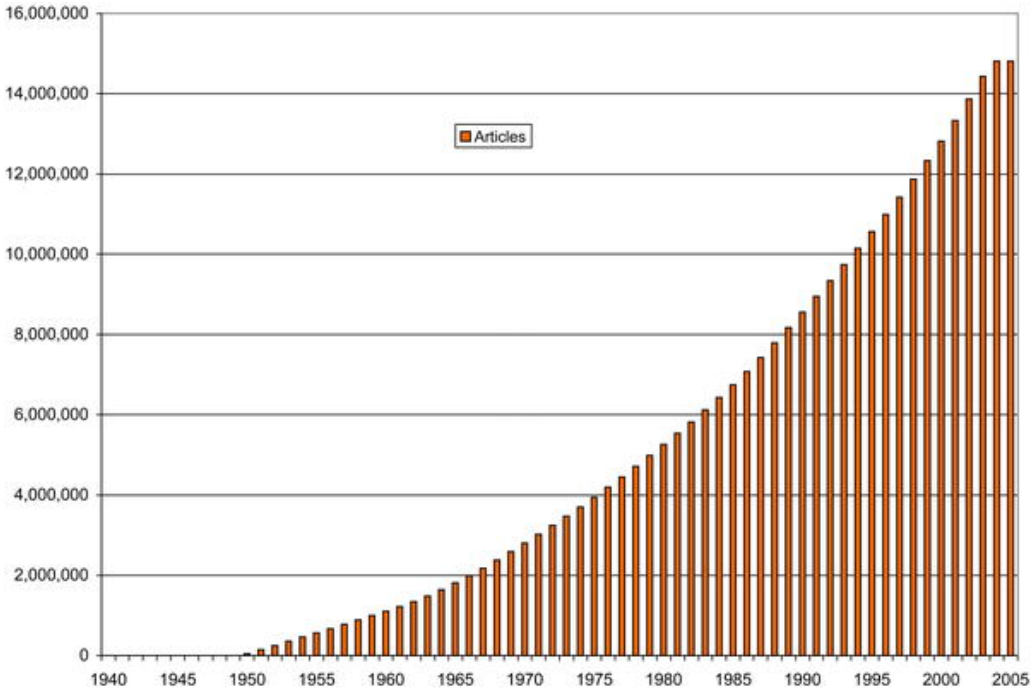

Figure 1.11 - Growth of MEDLINE: More than 14 million abstracts are now available from MEDLINE covering the articles published in last 65 years. More than 2 million abstracts were added to the database last year.

The work described in this thesis attempts to analyse the text from abstracts in MEDLINE as well as full text articles. It concentrates on the development of approaches for reliably extracting useful information in order to generate databases and carry out hypothesis testing. In the rest of this section I first provide an overview of the work on information extraction (IE) specifically that related to biology, and then I describe common methods utilized in the field and in this thesis.

1.2. - Automated handling of text

Automated handling of text is an active research area, spanning disciplines like Information Retrieval (IR), Text Categorization, Natural Language Processing (NLP), Information Extraction (IE) and Text Mining. IR mostly deals with finding relevant documents that satisfy a particular information need within a large database of documents. Thus, given a user query the IR system perform a query expansion and aims to provide all relevant documents in the database to the user. Text classification (or text categorization) is a

related problem. The aim of text classification is to automatically assign semantic categories to natural language text.

NLP is a broad discipline concerned with all aspects of automatically processing both written and spoken language. IE is a sub-field of NLP, centred on finding explicit entities and facts related to events or scenarios in unstructured texts. Finally, text mining is the combined, automated process of analysing unstructured, natural language text in order to discover information and knowledge that is typically difficult to retrieve. Most of the IE methods, whose goal is to extract descriptions of events, operate at the sentence level. The complete description of entire events is achieved by techniques in discourse resolution.

1.2.1. - Logical View of Documents

Modern computers are making it possible to represent a document by its full set of words and it is the most logical view of the documents. However, with very large number of documents, even modern computers have to reduce the set of representative keywords (Figure 1.21). This can be accomplished through removal of stopwords (such as articles and connectives), use of stemming (which reduces distinct words to their common grammatical root), and identification of noun groups (which eliminates adjectives, adverbs, and verbs). Further text operations including compression may be employed (see Appendix A). Text operations reduce the complexity of document representation and allow moving the logical view from that of a full text to a set of index terms or keywords. Keywords provide the most concise logical view of the document but its usage might lead to retrieval of poor quality.

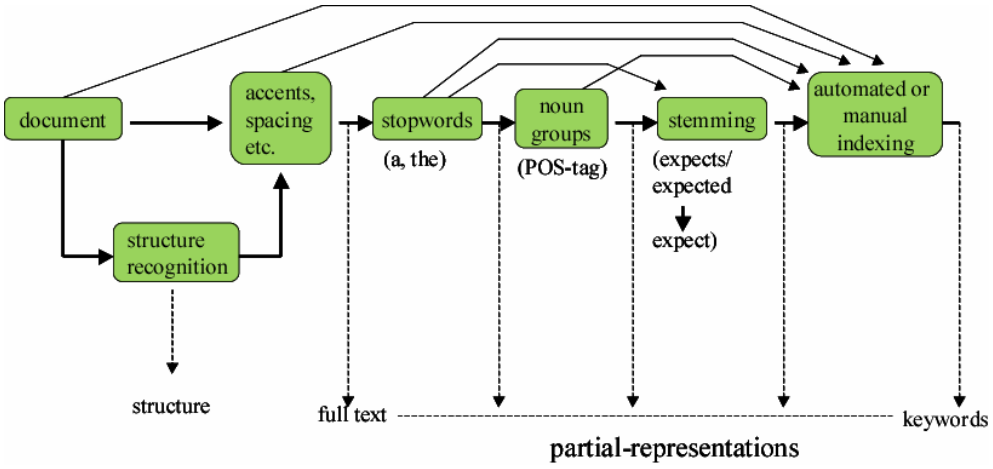

Figure 1.21 - Reducing documents to partial-representations. Documents could be reduced to partial representations using various text operations. Some of the operations include spacing, stop word removal, part of speech tagging, and stemming.

1.3. - An overview of IE methods

There are many ongoing efforts for Biomedical text mining, but the potential of these methods is yet unrealized. Text mining tools are not part of the standard arsenal of biomedical researchers the way search engines and sequence alignment tools are. However, just like the sequence analysis field, the bio-IE is borrowing tools developed by the computational linguists for general English. The IE efforts in biology involve named entity extraction, extraction of relationships and descriptions of entire events, hypothesis generation and design of general purpose IE methods. Most efforts in information extraction to date focus on using a curated lexica or natural language processing for identifying relevant phrases and facts in texts. The techniques include regular expression matching, co-occurrence of terms, statistical methods, advanced parsing and machine learning methods. Please see (Andrade and Bork, 2000; Cohen and Hersh, 2005; de Bruijn and Martin, 2002; Krallinger et al., 2005; Shatkay and Feldman, 2003) for in depth discussions of current approaches for IE in molecular biology.

1.3.1. - Named entity extraction

In order to identify and extract structured from information texts, recognizing entity names is the important first step. Other information can be mined subsequently from a text tagged with entity names. Such a task, called named entity (NE) recognition, has been well described in the IE Literature. In Message Understanding Conference (MUC), the task of named entity recognition is to recognize the names of persons, locations, organizations, etc. in the newswire domain.

The first challenge in biomedical information extraction is to recognize NE like gene or proteins GENIA ontology includes 23 distinct entities including multi-cell, mono-cell, virus, body-part, tissue, organism, cell-line and others (Kim et al., 2003). Relationship (e.g., protein-protein interactions, signal transduction pathways) extraction can be performed from texts tagged with entity names (Hoffmann et al., 2005). Various features of biomedical named entities could be used for accurately identifying them. Such features include word formation pattern, morphological cues, part of speech tags, head noun trigger and dependency relationships (Zhou et al., 2004). Several methods are now available for named entity extraction (Alphonse et al., 2004; Collier et al., 2002; Fukuda et al., 1998; Tanabe and Wilbur, 2002) and at least two community wide efforts for this task have been organized in the past . NE extraction wasn't attempted in this thesis instead the goal was higher level IE for event extraction.

1.3.2. - Relationship extraction

The earliest effort for relationship extraction from biomedical texts was aimed at extracting sentences discussing gene location on chromosomes using hidden Markov models (Leek, 1997). In the case of sentences describing location, the constituents are gene and chromosome names, words describing location, and terms denoting experimental methods that validate the location of a gene on a chromosome. Craven et al developed systems to distinguish fact-bearing sentences from "uninteresting" sentences for

identifying protein sub-cellular localization and gene-disorder association (Craven and Kumlien, 1999). Their naïve Bayes classifier that doesn't use grammatical rules achieved a precision of 77% and a recall of 30%. The classifier that used grammatical rules and parsing of sentences achieved a higher precision (92%) but a lower recall (21%). An important result of these experiments is the comparison of classifiers to a baseline method, which uses co-occurrence alone. The latter method decides that a sentence reports a "subcellular localization" fact if both a protein name and a localization word occur in it. This simple method, which is currently most popular in context of literature mining in Bioinformatics, reaches a much lower precision than the classifiers (about 35% precision at recall 30% and 45% precision at recall 21%). The co-occurrence based method can reach a higher level of recall (~70%) without losing much in precision (~40%). However, at this higher recall level, a naïve Bayes classifier with a noisy-or combination still reaches a somewhat higher level of precision (~45-50%). The study suggests that classifiers at the sentence level have the potential to improve precision of information extraction, in the biomedical context, over co-occurrence-based methods.

Co-occurrence based methods have been used widely for detecting protein-protein interactions while analysing gene expression data. The problem with the co-occurrence based methods is that it yields no information about the types relations described in the literature and therefore co-occurrence results may be misleading (e.g., in case of negative sentences). NLP based approaches have been carried out for extraction of protein-protein interactions, interesting keywords, gene-drug relationships, mutations, roles of residues in protein function, and regulation of gene expression. These methods utilize regular expressions, methods from the domain of machine learning, and computational linguistics for natural language processing (Alphonse et al., 2004; Blaschke et al., 1999; Donaldson et al., 2003; Marcotte et al., 2001; Novichkova et al., 2003; Ono et al., 2001; Pustejovsky et al., 2002; Rindflesch et al., 2000; Sekimizu et al., 1998).

Relationship extraction: an example

Saric and co-workers carried out relationship extraction for generating gene regulatory networks for the Baker's yeast (*S. Cerviciae)* from the information published in MEDLINE (Saric et al., 2004). Their work involved six levels that are tokenization and identification of multi-words, POS-Tagging, semantic labelling, named entity chunking, relation chunking, and output and visualization. In the process, they improved tagging accuracy of a POS tagger trained on general text by retraining it on GENIA corpus. They defined a simplified ontology that represented biological knowledge about transcription regulation (Figure 1. 31). They concentrated on sentences with different verbs that defined events of activation (e.g., enhance, increase, and induce), repression (e.g., blocks, decreases, down regulate), regulation (e.g., affect and control) and gene transcription (e.g. encode). Their system identified 441 pair wise relations from 58,664 abstracts with an accuracy of 83-90% (Figure 1.31).

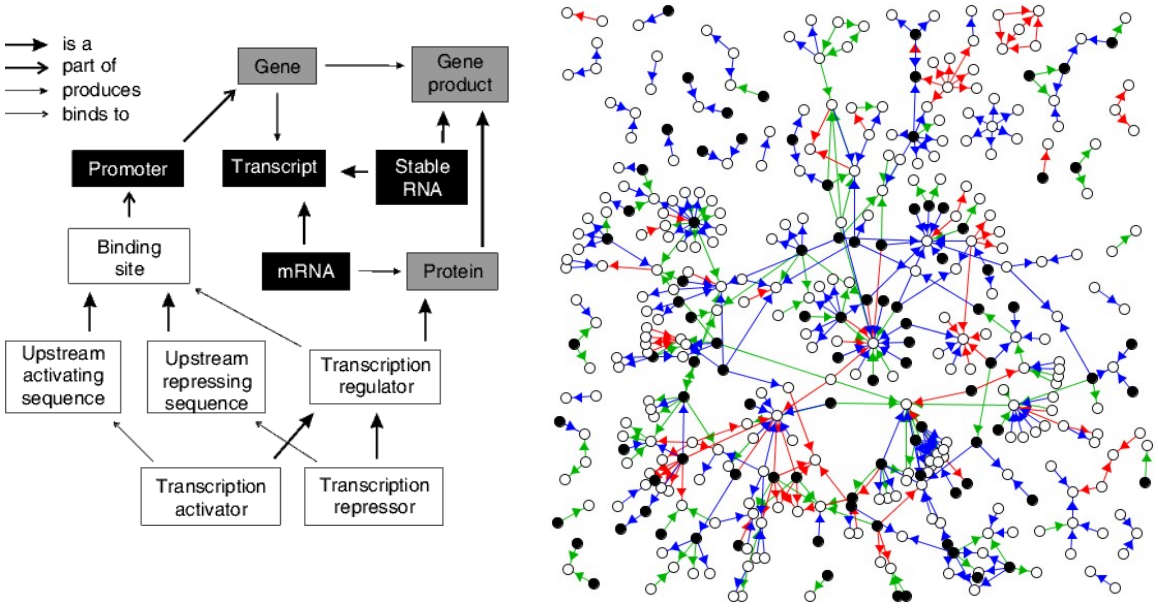

Figure 1.31 - Relationship extraction for transcription regulation: The first half of the figure shows a simplified ontology for transcription regulation process. The box colours for each term signifies its semantic role in relations: regulator (white), target (black), or either (grey). The extracted network is shown in the right side with the similar roles in red, blue and green colours, respectively (Saric et al., 2004).

1.3.3. - Hypothesis generation

Knowledge in biomedical literature has also been used for hypothesis generation. By combining knowledge in gene ontology and MeSH terms, the G2D system proposes candidate disease genes in human genome for genetically inherited diseases (Perez-Iratxeta et al., 2002). This was achieved by calculating fuzzy associations between different keyword systems in GO (Lewis, 2005), MEDLINE and LocusLink (Pruitt and Maglott, 2001). On a smaller scale, Srinivasan and Libbus identified therapeutic usage of turmeric on the retinal diseases, Crohn's disease and spinal cord injuries (Srinivasan and Libbus, 2004).

1.3.4. - Integration Frameworks

Several research groups are developing integrated text mining frameworks intended to be able to address a variety of user needs. The MedScan system combines lexicons with syntactic and semantic templates into extract relationships between biomedical entities (Daraselia et al., 2004; Novichkova et al., 2003). The PubMatrix tool displays two dimensional comparisons of gene names and functional terms based on combining the results of multiple queries to PubMed (Becker et al., 2003). The BioRAT system is another template based system that combines a template design tool with a web spider that locates and retrieves full text journal articles (Corney et al., 2004). The Textpresso system uses a specially created ontology to flexibly combine keywords and concept co-occurrence searching of *Caenorhabditis elegans* literature at the sentence level (Muller et al., 2004). It performs both the IR and the IE tasks. There are

other frameworks such as TXTGate (Glenisson et al., 2004) and those reported by Nenadic (Nenadic et al., 2003) and Chiang (Chiang et al., 2004). All the systems are still in their development phase and their value haven't been strictly assessed. Representative systems in text-mini are summarized in Table 1.31.

1.3.5. - Ontologies in Biology

The term ontology is becoming more and more popular in biology and related fields. It is used to refer too many things, amongst which are controlled vocabularies (e.g., Medical Subject Heading Terms), taxonomies (e.g., Gene Ontology), conceptual model of a given domain (as description of rules to infer new knowledge), or a combination of parts of or all of the above. Every term in ontology could be considered as representing a concept. Here I describe two ontologies commonly used in molecular biology research.

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the National Library of Medicine and used for indexing, cataloguing and searching for biomedical and health related information and documents. One or more MeSH terms, comprising one or more concepts, grouped together, form a descriptor class. The descriptor class is a basic building block of the thesaurus. Hierarchical relationships are defined (as parent-child) among the descriptor classes. Each abstract available from PubMed database at NCBI are annotated with up to 20 MeSH terms. User queries are expanded and searched against the MeSH terms. The hierarchical structure in MeSH allows for retrieval of broader or narrower retrieval sets.

The gene ontology (GO) project provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology (Lewis, 2005). GO ontologies describe attributes of gene products in terms of molecular function, biological process and cellular component. The vocabularies are structured in a classification that supports 'is-a' and 'part-of' relationships. For many purposes, in particular reporting results of GO annotations of genome, cDNA collections and microarray data the curators provide 'GO slim', a subset of GO. Recently, Sequence Ontology (SO) has been made available to describe the classification and standard representation of sequence features (Eilbeck et al., 2005).

Task 2 of BioCreative 2004 workshop was focused on extracting relevant GO codes from the free text. Full-text articles were used in the task and the participating systems were evaluated by mouse genome curators for their usefulness. The system precisions ranged between 2% and 80%. The recall wasn't evaluated. This task is particularly difficult as the systems had to get the text, the gene and the GO codes all correct simultaneously.

	Basic features	URL
Repositories		
PubMed/Entrez	Biomedical citation retrieval	www.ncbi.nlm.nih.gov/entrez
GENIA Corpus	Annotated corpus related to human blood cell transcription factors	www-tsujii.is.u-tokyo.ac.jp/GENIA
BioCreative corpus	Corpus of protein annotation of relevant text pages	www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
Assessments		
BioCreative Challenge	Text mining of protein names and annotations	www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
KDD challenge	Information extraction of Drosophila gene expression information	www.biostat.wisc.edu/~craven/kddcup/tasks.html
TREC Genomics track	IR, document classification and question answering in biology domain	lr.ohsu.edu/genomics
NLPBA challenge	Protein and gene name identification	www.genesis.ch/~natlang/JNLPBA04
Information Retrieval		
PubMed/Entrez	Biomedical literature retrieval tool	www.ncbi.nlm.nih.gov/entrez
XplorMed	Iterative retrieval and extraction of abstracts	www.bork.embl.de/xplormed/
Google Scholar	Literature search engine	scholar.google.com
CrossRef search	Full content search engine	www.crossref.org/crossrefsearch.html
Name recognition		
AbGene	Protein/gene name tagger	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe
GAPSCORE	Protein/gene name tagger	bionlp.stanford.edu/gapscore
NLProt	Protein/gene name tagger	cubic.bioc.columbia.edu/services/nlprot
Protein (set) function		
PubGene	Text mining tool for microarrays	www.pubgene.org
MedMiner	Extract gene relevant sentences	discover.nci.nih.gov/textmining/main.jsp
iProLINK	Protein annotation and tagging	pir.georgetown.edu/prolink
Textpresso	C.elegans literature IR/IE tool	medblast.sibsnet.org
KAT	Annotate protein from scientific literature	www.bork.embl.de/kat
Protein Interactions		
Chillbot	Relationship extraction tool	www.chillbot.net
GeneScene	IE of regulatory pathways	econport.arizona.edu:8080/NetVIs/index.html
PreBIND	Classifier of protein interaction documents	bind.ca
Protein network exploration		
iHOP	Literature based gene and protein network	www.pdg.cnb.uam.es/UniPub/iHOP
STRING	Gene and protein network, uses literature	string.embl.de
Knowledge discovery		
ARROWSMITH	Extended MEDLINE search	kiwi.uchicago.edu
G2D	Identification of disease genes	www.bork.embl.de/G2D
BITOLA	Literature based biomedical discovery	www.mf.uni-lj.si/biotla

Table 1.31 - Representative list of systems for biomedical text handling: [modified from (Hoffmann et al., 2005)]

1.4. – Event extraction

Event extraction is defined as automatically finding, within a text, instances of a specified type of event, and filling a database with information about the participants and circumstances of the event.

1.4.1. - Events in Molecular Biology

According to Gene Ontology (GO) Consortium the term *biological process* refers to a broad category of biological tasks accomplished via one or more ordered assemblies of molecular entities or gene products (Lewis, 2005). It often involves transformation, in the sense that something goes into a process and something different comes out of it. Examples of biological processes are cell growth and maintenance, signal transduction, metabolism and biosynthesis etc.

A biological process can be subdivided into temporal and spatial molecular events. Each molecular event is carried out by a gene product or well-defined assemblies of them. For example, *phosphorylation* of a protein molecule by a protein kinase is a molecular event, which is a part of the cellular signaling process. Similarly, *transcription* of a gene by a polymerase is a part of the gene expression process. Hence, by definition, a molecular event or a disruption of it will have a local effect in terms of the process that it is a part of and an observable or phenotypic effect in terms of the overall effect of disruption of the entire process. For example, pre-mature termination of translation due a pre-mature stop codon arising by a *mutation* in the coding region of a gene could be considered as a local effect and the disease state of an organism due to deficiency of that protein could be considered as the phenotypic effect.

1.4.2. - Template based extraction of relationships and events

IE, based on the MUC tradition of task segmentation (McCallum and Nigam, 1998) works fundamentally by using predefined frames and slots in agreement with a specific scenario describing user requirements. Such systems typically use regular expressions to match facts for the event to be extracted in each sentence. Each logical form is based upon the syntactic relationship between components in each sentence. For instance, if we wanted to extract facts relating to a scenario (*alternative splicing*) then patterns such as "np (the difference) + in (in) + np (the splicing pattern) + pp (in (of) + np (the first intron)) + vp (is mouse-specific)" and "np (Alternative splicing) + vp (vbz (occurs) + in (at) + np (the 5' region) + pp (of the ATR1))" could be developed as templates. Sentences constituents which contain information about the mechanism, exon/intron position, gene names, species etc. need to be extracted. This task is difficult because a single event can nearly always be written in a variety of syntactic forms due to mechanistic and linguistic aliterations.

1.4.3. - Spray alterations and problem of syntactic patterns in IE

The following simple example involves a linguistic phenomenon sometimes called locative alternation or *spray alternation* (Levin, 1993). The verb *spray* may express its arguments in at least two different ways, i.e. (a) "*Peter sprayed water on his flowers.*" and (b) "*Peter sprayed his flowers with*

water." Thus, two syntax-based regular expressions plus some information about NE as np (peter) + vp (spray) + np (object1) + pp (on) + np (object2) and np (peter) + vp (spray) + np (object2) + pp (with) + np (object1) are required.

Extraction patterns can be hand built or based on machine learning (ML) from a corpus (a sample of annotated text) or from a few patterns which are known to be good indicators of the topic of interest (seed patterns) to reduce the cost and time in constructing patterns manually (Alphonse et al., 2004; Hobbs et al., 1997; Riloff, 1993; Riloff, 1996; Yangarber, 2003; Yangarber et al., 2000). However, to extract the relations between objects in the complex sentences that occur in technical and scientific texts requires deeper knowledge. Most of the existing systems use a set of rules relevant to syntactic roles (e.g. subject, object, and modifier) obtained from parsers, together with surface level patterns to extract the interactions between genes or gene products from the biological literature (Ono et al., 2001; Pustejovsky et al., 2002; Rindflesch et al., 2000; Sekimizu et al., 1998). Although extending the systems with syntactic roles or syntactic functions can help achieve better performance compared to the pure pattern-matching approach, errors resulting from a lack of semantic understanding still remain (Ono et al., 2001). For example, the system of Ono et al. will incorrectly extract a protein interaction between "*Msp1p*" and "*Dec1p*" from a sentence "*These findings suggest that Msp1p is a component of the secretary vesicle docking complex whose function is closely associated with that of Dec1p.*", because it conforms to the pattern "A [associate with] B" predefined within the system. In this respect deeper knowledge, describing the semantic relationships between verbs and its arguments is needed to overcome difficulties posed by syntactic patterns.

1.4.4. - Need for Semantic Relationships in Molecular Event Extraction

The issue of syntactic patterns is encountered in texts of all domains including scientific texts (Figure 1.41).

(1) [$_{A}$ One mutation] eliminates [$_{B}$ the BamH1 site] in [$_{C}$ exon 7] and …
(2) The same high level of activation of B-Raf occurs only when [$_{B}$ all three sites] are eliminated
(3) One of the three remaining families carried [$_{A}$ a 3-bp in-frame deletion] that would eliminate an [$_{B}$ asparagin residue] within [$_{C}$ a kinase domain of the product]; the …

Figure 1.41 - Example of different forms of eliminate: Three different sentences containing the predicate eliminate illustrate the existence of multiple syntactic patterns in the biomedical texts. These sentence may be written in various forms but convey the information marked as […$_{A}$] or […$_{B}$] or […$_{C}$].

The sentences show different instances of the event *eliminate* taken from corpus of biomedical texts (Figure 1.41). Here, there may be 3 different pieces of information to be extracted, i.e. A – causal agent of the event, B – the entity being removed, C – location at molecular (sequence) or cellular level where the entity is being removed. Sentence 1 shows the simple indicative form for which the syntactic extraction pattern would be "A eliminates B in C" (where A=*One mutation*, B=*the BamHI site* and

C=*exon7*); sentence 2 shows the passive form, without mention of A and C, for which the syntactic extraction pattern would be "B are eliminated" (where B=*all three sites*); sentence 3 utilizes a different preposition compared to sentence 1 in order to mention C, for which syntactic extraction pattern would be "A would eliminate B within C" (where A=*a 3-bp in-frame deletion*, B=*an asparagines residue* and C=*a kinase domain of the product*).

(1) Northern blot analysis with mRNA from eight different human tissues demonstrated that [$_{A}$ the enzyme] was expressed exclusively in [$_{C}$ the brain], with [$_{B}$ two mRNA isoforms of 2.4 and 4.0 kb].
(2) [$_{A}$ Two equally abundant mRNAs for IL8RA] , [$_{B}$ 2.0 and 2.4 kilobases in length] , are expressed in [$_{C}$ neutrophiles] and arise from usage of two alternative polyadenylation signals.
(3) This "functional allelic exclusion" is apparently due to control of the TCR assembly process because these [T-cells] express [$_{A}$ RNA and protein for all four transgenic TCR proteins].

Figure 1.42 - Example of different forms of *express*: The surface variation of linguistic expressions is clear from sentences (1)-(3) for the event *express*. Sentence (3) emphasizes the fact that domain knowledge is necessary for understanding the sentence (see the text).

In the sentences describing the event *express* (Figure 1.42) the information slots are A – expressed entity, B – physical property of the expressed entity, and C – location referring to the organelle, cell or tissue. In sentence 1, (where A= *the enzyme*, B=*two mRNA isoforms of 2.4 and 4.0kb*, C=*brain*) the information needed to describe the event with respect to the slot B is marked by using a prepositional phrase, but using an appositive form in sentence 2, (where A=*two equally abundant mRNAs for il8ra*, B=*2.0 and 2.4 kilobases in length*, C=*neutrophils*), seemingly not playing an important role in the description of the event in which it participates. Sentence 3, (where A=*RNA and protein for all four transgenic TCR proteins* and C=*T cells*, without mentioning B) illustrates a different problem involving "T-cells", because from a biological perspective "T cells" would qualify as source/location rather than as an agent from a linguistic point of view.

These examples show that extraction using regular expressions around syntactic information of the surface texts would not be adequate for high performance IE due to complexities in surface structures. Instead, mapping of various surface structures into the same predicate argument structures (introduced below) would be beneficial, as it represents the information describing the arguments and the semantic roles these arguments play with a verb that conveys a certain event.

1.5. - Predicate Argument Structures

In natural language sentences, an event or relation is expressed as a verb, and the participants involved are expressed as the arguments of the verb. A verb, which indicates a particular type of event conveyed by a sentence, can exist in its verbal form, its participial modifier format or its nominal form. For example, the normal form of a verb used to describe the event "making something active" would be *activate*, its participial modifier format would be *activating* or *activated*, and its nominal format would be *activation*.
The participants in the description of events or relations may have specific roles. The common technical name for them is ***argument structure*** and the verb that specifies the argument structure is called the ***predicate***. It is common to refer to the contents of arguments with labels where such roles are usually specified. Meaning can be determined in several ways such as a domain or predicate-specific fashion such as *catalyst* and *reaction being catalyzed* in case of the first and second arguments to the predicate *catalyze*. Alternatively, functional roles can be employed such as thematic relations that try to express some linguistically motivated aspect of the argument's behaviour such as *agent*, *location* or *experiencer*.

Traditional IE systems that use regular expressions based on shallow chunking at the phrase level (e.g. noun phrase, verb phrase, preposition phrase) capture weak notions of 'argument' for event predicates and their linear precedence. Such approaches seem to be inadequate to the goal of achieving high completeness and accuracy in event extraction. In recognition of this several major projects for generating predicate argument structures (PAS) have now begun for general English from newswire texts (Baker et al., 1998; Kingsbury and Palmer, 2002; Kingsbury et al., 2002; Kipper et al., 2000). They examine relations that exist between the constituents in a sentence with the key assumption that those arguments correspond to major objects in events of interest. Although constructing PAS frames by hands seems to be expensive in terms of time and effort, particularly where this requires insights from domain specialists, this is justified as they provide a systematic reference guide for improving performance compared to the ad-hoc pattern building.

For PAS to be realized within IE, three major knowledge components are required: (1) a hierarchy of concept categories for the objects of interest; (2) a definition of predicate-argument frames and the semantics of their arguments; and (3) the mapping rules that define how to transform the relevant parts of a surface sentence to the arguments in the PAS frame. Currently (1) is already quite advanced with several controlled vocabularies such as MeSH (Nelson et al., 2000) or Gene Ontology (Lewis, 2005) are now widely in use. At a more modest level core domain specific ontologies for individual annotation schemes such as the GENIA project (Kim et al., 2003) have also been proposed. However, there are no proposals for (2) for biomedical texts which may serve as the basis on which annotated resources could be developed for machine learning approaches to (3).

1.5.1. - Resources for PAS

Several major projects provide PAS for verbs in common English. These projects include VerbNet (Kipper et al., 2000), FrameNet (Baker et al., 1998), and PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002). However, there are methodological differences among these three projects. Example PAS of verbs *sell* and *rent* from these projects illustrates this point (Figure 1.51).

VerbNet : PAS for verb group: Give
Verb Members: give, sell, rent, render, refund, peddle, pass, loan, lend, lease
Arguments: 0 : Agent 1: theme 2: recipient
Sentence 1: [$_{Arg0}$Michael] sold [$_{Arg1}$it] for $60 a bottle.
Sentence 2: [$_{Arg0}$Mary] rented [$_{Arg1}$a room] to [$_{Arg2}$John] for a week, then evicted him

FrameNet : PAS for Event: Commerce_sell
Event Definition: Basic commercial transaction from the perspective of the seller
Verb Members: sell, rent, charge, lease, retail, vend
Arguments: 0 : seller 1: goods
Sentence 1: [$_{Arg0}$Michael] sold [$_{Arg1}$it] for $60 a bottle.
Sentence 2: [$_{Arg0}$Mary] rented [$_{Arg1}$a room] to John for a week, then evicted him

PropBank	
Verb: Sell Arguments: 0: seller 1: thing sold 2: buyer 3: price paid 4: term	Verb: Rent Arguments: 0: landlord 1: thing rented 2: renter 3: rent 4: term
Verb Members: sell, rent, charge, lease, retail, vend	
Sentence 1: [$_{Arg0}$Michael] sold [$_{Arg1}$it] for [$_{Arg3}$$60 a bottle].	
Sentence 2: [$_{Arg0}$Mary] rented [$_{Arg1}$a room] to [$_{Arg2}$John] for [$_{Arg4}$a week], then evicted him	

Figure 1.51 - PAS definitions for sell and rent as defined by PropBank, VerbNet and FrameNet.

PropBank defines two distinct PAS for two distinct verbs while there is a single structure for both verbs in the case of VerbNet and FrameNet (Figure 1.51; top panel). In VerbNet, a general PAS is defined for a group of verbs that share similar syntactic behaviour, as suggested by Levin's alternations theory (Levin, 1993). Thus, VerbNet has a common PAS frame for *give, sell* and *rent.* Argument roles for all of these verbs are assigned for *agent*, *theme*, and *recipient* illustrated by example sentences 1 and 2. In the case of FrameNet, PAS is defined based on the underlying principal of what users or applications expect to see for a specific event definition.

FrameNet's PAS for event *Commerce_sell* expects only argument *seller* and *goods* from the event driven by any verb in a set of verb members (Figure 1.51; middle panel). Considering the annotation on sentence 1 in all projects, "All Brownstein" is annotated as *seller*, *agent*, and *seller* in PropBank, VerbNet, and FrameNet respectively. Similarly, there is also an argument to support the annotation of "it" in all projects. But, only the PropBank scheme has an argument labeled *price paid* to support element "$60 a bottle" of sentence 1 which is an important participant of the event describing the selling activity. Moreover, a constituent "a week" in sentence 2 is considered to be an argument labeled as *term* only by the PropBank scheme. The arguments like *price paid* for the events driven by a verb *sell,* and an argument *term* for events driven by a verb *rent,* are considerably important for down stream user applications. Also, in contrast to VerbNet and FrameNet, PropBank defines individual verb-specific PAS frames which are likely to contain more detailed specifications of arguments than are possible for verb groupings (Figure 1.51; bottom panel). Moreover, PAS construction in a more verb-specific manner than either VerbNet or FrameNet would assist explicitly in discovering rules for mapping from surface syntactic structures to underlying semantic propositions. Hence, PropBank's scheme for defining PAS is desirable as a basic starting point for generating PAS frames in molecular biology.

1.5.2. - Introduction to PropBank

PropBank PAS frames are based on an analysis of sentences in the Wall Street Journal corpus. In PropBank a verb may get more than one PAS frames if the verb sense and its argument set differ, underlying the fundamental assumption that syntactic frames are direct reflections of underlying semantics. For example, PropBank defines following three distinctive PAS frames for the verb *run* on account of sense variation (Figure 1.52). Each structure contains its own set of arguments labelled with semantic roles. A semantic role of an argument represents a semantic relationship between the argument and its related verb. Though, not all arguments of a given verb may be present in a given sentence. The example sentence in Figure 7 (left panel) illustrates this point i.e. only *Arg0* and *Arg1* occur in this sentence without the occurrence of *Arg2*, *Arg3*, and *Arg4* though all arguments are defined as core arguments of the PAS. In each PAS, arguments are labelled ranging from *Arg0* up to *Arg5* with a mnemonic label indicating its predicate-dependent role.

PAS for Verb: RUN

Sense: operate, proceed

Arguments:

Arg0: operator

Arg1: machine, operation, procedure

Arg2: employer

Arg3: coworker

Arg4: instrumental

Example:

Mr. Stromach wants to resume a more influential role in running the company.

Arg0: Mr. Stromach

REL: running

Arg1: the company

PAS for Verb: RUN

Sense: walk quickly

Arguments:

Arg0: runner

Arg1: course, race, distance

Example:

John ran the Boston Marathon.

Arg0: John

REL: ran

Arg1: the Boston Marathon

PAS for Verb: RUN

Sense: encounter

Arguments:

Arg0: encounterer

Arg1: thing encountered

Example:

John ran into problems with his dissertation. Again. And again.

Arg0: John

REL: ran

Arg1: problems with his dissertation

Figure 1.52 – Three distinct PAS definitions for the verb run as defined in PropBank: PropBank defines different predicate-argument structures on account of verb sense variation (Kingsbury and Palmer, 2002; Kingsbury et al., 2002). Thus, three distinctive predicate-argument structures are defined for the verb run. PAS for each sense contains its own set of arguments labeled with semantic roles.

Besides these core arguments defined in PAS, some arguments known as adjuncts are traditionally not defined in PAS because they are linguistically not required to minimally define the event. PropBank does consider adjuncts when annotating sentences, and provides labels such as ArgM plus tags such as TMP for temporal information, LOC for locative information, PRP for a reason or motivation, etc. More information on the PropBank project could be found in (Kingsbury and Palmer, 2002; Kingsbury et al., 2002). After manually defining PAS, PropBank has annotated the Wall Street Journal corpus, which already contains constituency and dependency information from the TreeBank project (Marcus, 1994).

1.6. - Classification using inductive machine learning

The goal of learning inductive classification is to infer a classification rule from a sample of labelled training examples so that the learner classifies new examples with high accuracy (see Appendix A). A proper definition of inductive learning is given in the Methods section. Learning algorithms like Perceptron, Winnow and support vector machines determine a linear decision boundary (hyperplane) for the binary classification of the data. However, in case of data that are not linearly separable, kernel methods can be used to transform them to linearly separable form. In the context of NLP, machine learning

algorithms could be used for text/sentence classification or for giving appropriate semantic labels to various arguments in PAS. Details of learning algorithms used in this work can be found in appendix B.

1.7 - Generation of Alternative transcripts

Estimates for the total number of human and mouse protein-coding genes currently fall in the range of 20-25,000, whereas those of simpler organisms like Drosophila melanogaster (fruit fly) and Caenorhabditis elegans (worm) are lower with ~13,000 and 18,000 genes, respectively. Given that the mammalian genomes have less than twice as many genes as the fruit fly and the worm, it is generally believed that the phenotypic complexity of higher organisms is achieved not only by higher gene numbers, but also by multiple mRNA transcript isoforms (Landry et al., 2003).

Results from Human genome tiling arrays suggests that the number of transcripts encoded by the genome is at least 10-fold greater than the number of genes (Bertone et al., 2004; Cheng et al., 2005; Kampa et al., 2004). Similarly, the recent analysis of FANTOM consortium data suggest at least one order of magnitude more transcripts than estimated 22,000 genes in the mouse genome (Waterston et al., 2002). The generation of multiple alternative transcripts is important for the complexity and evolution of eukaryotic organisms (Boue et al., 2003). In addition, their spatial and temporal expression patterns are believed to be one of the important factors behind the functional specificity of different tissues and organs. Moreover, defects in these processes are associated with various diseases (Garcia-Blanco et al., 2004). Thus, developing an exhaustive catalogue of alternative transcripts is a crucial task in order to fully understand the complexity of eukaryotes (Graveley, 2002).

High-throughput experiments and computational analyses dominate the mapping of the alternative transcript universe. However, the quality and the biological meaning of these assignments should be assessed against a highly reliable benchmark set, which can be extracted from single-gene studies published in the scientific literature. In addition, computational tools to explore the evolutionary conservation of mechanisms that generate transcript diversity are under development, which will also require a trustworthy set for rule learning. Synergy in using these mechanisms and preference of usage for various mechanisms by different tissues/organ systems is also being explored. Manual curation of experimentally determined biological events (physical interactions, AS, disease phenotypes, etc.) to generate trustworthy knowledge bases is slow compared to the rapid increase in the body of knowledge represented in the literature. Hence, IE approaches developed in this thesis were utilized to extract information about mechanisms for generating alternative transcripts in eukaryotes. It is described in the later part of the thesis. Here a brief introduction to various mechanisms to generate alternative transcripts is given.

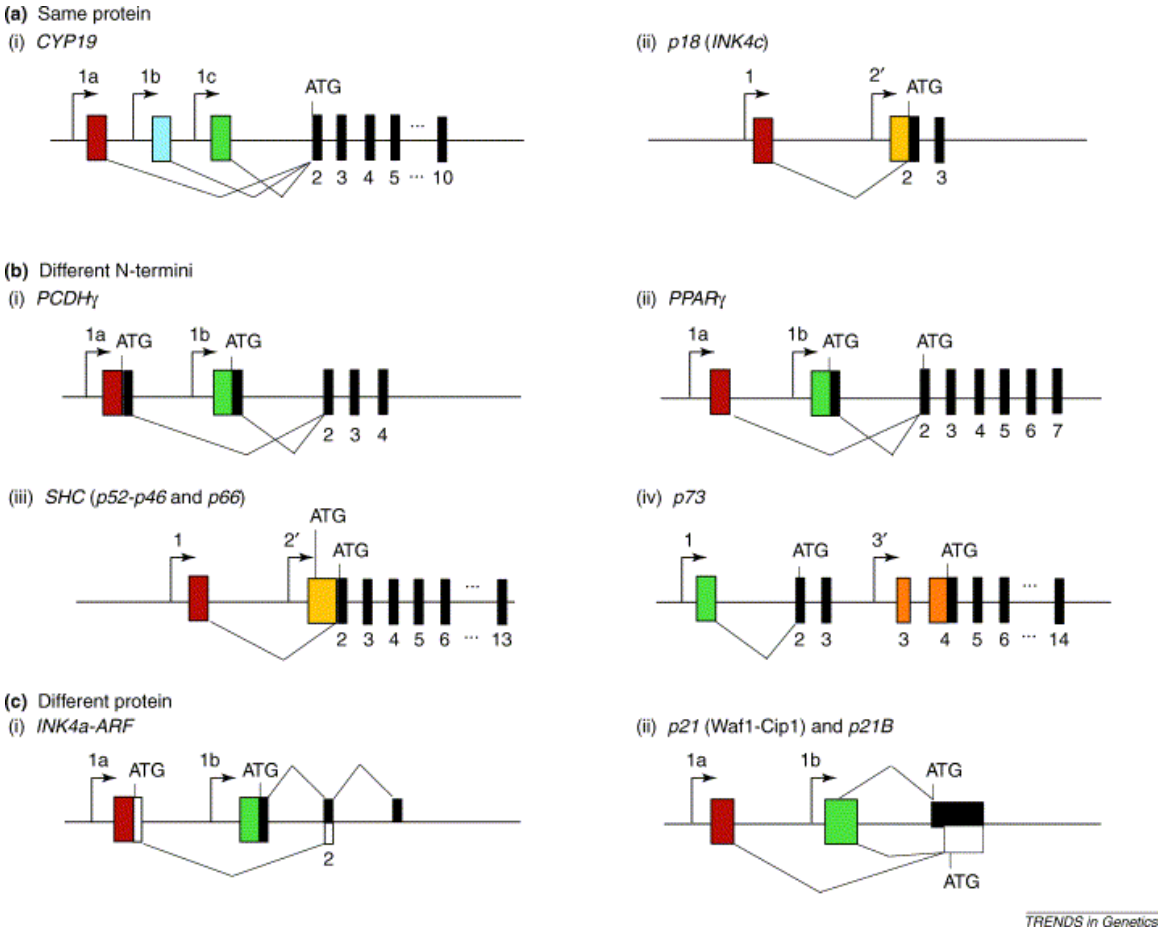

Figure 1.71 - Types and consequences of alternative promoters: (a) The use of alternative promoters (represented by arrows) does not result in protein isoforms because the variant 5′ initial exons (coloured boxes) are joined to a common second exon that contains the translation initiation site, shown as ATG. The black boxes illustrate coding exons and 3′ untranslated regions (UTRs) are not shown. Note that splicing, represented by solid lines, is only shown between the first and second exons for (a) and (b). (b) Using multiple promoters produces mRNAs that encode protein isoforms differing in their N-termini. (c) The use of the alternative promoters creates transcripts that code for different proteins as they are translated in different reading frames (represented by the black and white boxes). An example of a gene representing each type of alternative promoter usage present in both human and mouse is given with the exception of 1c, ii where the alternative promoter has only been identified in human. In some cases, not all promoters are shown.

1.7.1. - Alternative promoters

Alternative transcripts are generated using different mechanisms presumably working in concert. Alternative promoters (Figure 1.71) that are active in different tissues or at different developmental stages often regulate the expression of different mRNA isoforms, either directly through different transcription start sites or indirectly by promoter-directed exon inclusion in concert with alternative splicing (AS). For many genes for which multiple promoters have been documented, no variation in the resulting proteins has been reported. In these genes, although the mRNAs have alternative first exons, a common downstream exon contain the translation initiation site and therefore have the same open reading frame (Figure 1.71). Although no protein isoforms are generated in these instances, the mRNA variants differ in their transcriptional patterns and translation efficiencies. A recent analysis increased the estimate for alternative promoter usage by protein coding transcription units from 18-20% (Landry et al., 2003) to 58% in mouse genome (Carninci et al., PLoS Genet.; Submitted).

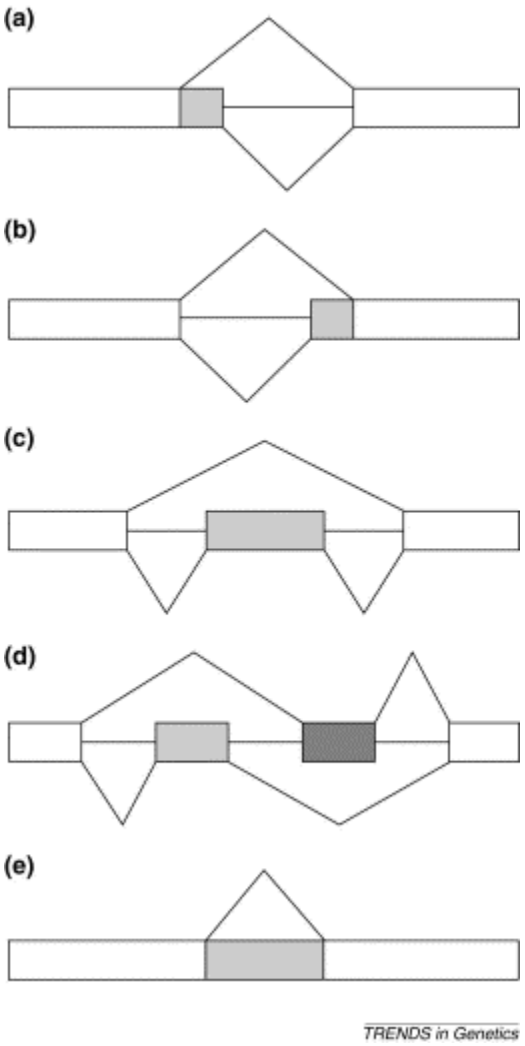

Figure 1.72 - Different mechanisms of alternative splicing: Every conceivable pattern of alternative splicing is found in Nature. Exons can have multiple 5′- or 3′-splice sites that are alternatively used (a, b). Cassette exons are fully contained exons that are alternatively used. Single cassette exons can reside between two constitutive exons such that the alternative exon is either included or skipped (c). Alternatively, multiple cassette exons can reside between two constitutive exons such that the splicing machinery must choose between them (d). Finally, introns can be retained in the mRNA and become translated (e). The constitutive exons are depicted as open boxes and alternative exons are shaded. The lines above and below the boxes show possible alternative splicing events.

1.7.2. - Alternative splicing

Various AS mechanisms are known: alternative 5′ or 3′ sites can result in exons of different size, exons can be included or skipped, or an entire intron may be retained (Figure 1.72). A total of 60% and more than 74% of multi-exon human genes are believed to undergo alternative splicing (Johnson et al., 2003; Kornblihtt, 2005). AS regulation not only depends on the interaction of splicing factors with their target sequences in the pre-mRNA but is coupled to transcription. AS could be differentially regulated and generate a repertoire of protein-protein interactions due to difference in protein isoforms.

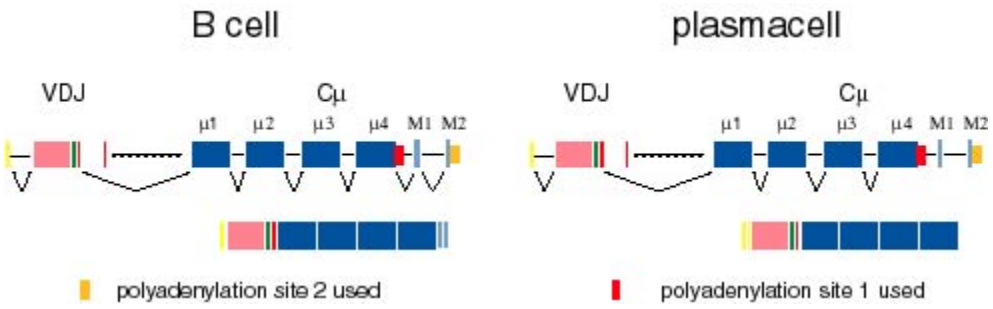

Figure 1.73: alternative polyadenylation for tissue-specific transcripts: Tissue-specific usage of alternative polyadenylation signal is illustrated here. It accompanies selection of a different 3' exon, resulting in different mRNA isoforms.

1.7.3. - Alternative polyadenylation

Alternative polyadenylation (AP), either alone or coupled with AS of 3′ terminal exons, may also generate transcript isoforms that are tissue- or developmental-stage-specific (Figure 1.73). Recent analysis based on EST data estimate at least 49% (human), 31% (mouse), and 28% (rat) of polyadenylated transcription units to utilize alternative polyadenylation (Yan and Marr, 2005). A portion of these alternative polyadenylation events results in new protein isoforms. The estimates of percentage of mammalian genes using these mechanisms are increasing with time.

II. Objectives

The major objectives of this work were:

1. An analysis of full text articles, in order to identify information rich parts of scientific articles and analyse distribution of different kinds of information in the different sections of research articles.

- The full text of an article contains more information than its Abstract. However, in approaching full text analysis, problems including those of storage, computational capacity, quality and organization of the information carried by different sections, context dependency etc. must be tackled.

2. Development of a database of predicate argument structures (PASBio) for verbs common to biomedical texts.

- A semantic lexicon is an essential module for any general purpose IE system. PASBio will function as a knowledge-base for reliable information extraction from free-form biomedical text to structured databases.

3. Application of NLP and machine learning tools like support vector machines for identifying sentences describing the generation of alternative transcripts, from MEDLINE abstracts.

- Corpus based inductive machine learning of patterns is a superior approach to writing rules for identifying text at the document, paragraph or sentence level.

4. Application of various NLP tools to generate a database of alternative transcripts.

- Generation of alternative transcripts is considered to be a major reason behind the phenotypic variation and evolution of eukaryotes. Hence, semi-automatic extraction of this information will be useful for communities interested in studying these events either experimentally or computationally.

- It also allows automated gene annotations assisting the work of database curators.

5. Analysis of knowledge from the database of alternative transcripts and its comparison with EST data.

- Here, the hypothesis that the data extracted from text can be used not only for assisting other methods, but also as a stand-alone source for testing new hypothesis and deriving conclusions, is explored.

III. Methods

3.1. - Analysis of full-text articles for comparison of information in different sections

3.1.1. - Text Corpus for the analysis of full-text articles

The aim of the analysis was to compare the information carried by different sections of a paper, especially the differences between the Abstract and the rest. Therefore, a set of full-text articles, with a regular section structure, namely having a defined Abstract, Introduction, Methods, Results, and Discussion (A, I, M, R, D) sections was used for the study. Another requirement was a certain homogeneity of style across the articles (for example, a similar length of the Methods section) and, since there is a great interest in the field of data mining on the detection of gene names, the subject should be related to Genetics. Thus, 104 articles published in *Nature Genetics* from June 1998 (volume 19, issue 2) to June 2001 (volume 28, issue 2), which comply with the AIMRD structure were chosen. Note that other journals, or even the Letters of the very same *Nature Genetics*, might have a different structure (for example, lacking separated I, M, R, D sections).

3.1.2. - Derivation of associations between the words of a section

Given a section from an article, the text was split into sentences using TreeTagger (Schmid, 1994), a standard part of speech tagger. The associations were computed between the words that are tagged as nouns as their part of speech category. Following (Perez-Iratxeta et al., 2002), the association between two words (w_i, w_j) can be modelled as the degree of inclusion of one word into the other (I_w)which can defined as the fuzzy binary relation given by: $\mu_{I_w}(w_i, w_j) = \frac{|\, w_i \cap w_j \,|}{w_i}$, that is, the ratio of the number of sentences where both words w_i and w_j co-occur to the number of sentences the word w_i occurs.

3.1.3. - Selection of Keywords

The work was aimed to compare the information carried by different sections of a paper, especially the differences between the Abstract and the rest. The work focused on the extraction of relevant words (keywords) regarding objects, detected as nouns from natural text by Tree-tagger (Schmid, 1994). It has been previously observed that words associated strongly to many other words are relevant to the matter that is dealt in the article (Perez-Iratxeta et al., 2001). Thus, in order to derive keywords from the section of an article, associations between the words in a particular section were computed. Here, sentences were taken as the unit of text to look for associations, that is, two words are associated in the context of a section

if they co-occur repeatedly in sentences within that section. A scoring scheme was devised that gave a score (K) that is higher for words with many and strong relations to other words. This measure was used to select words as keywords, in this case, related to objects such as proteins, genes, organisms, etc.

A word is considered as relevant if it establishes many and strong relations to other words for the text analyzed (Perez-Iratxeta et al., 2002). Therefore, in a given section, a score for a word wi is defined that is equal to $K_i = \sum_{j \neq i} \mu_{I_w}(w_i, w_j)$, normalized to the maximum value found for K of any word in that section. Then, the keywords of the section are defined as those words that have a K score above a certain value.

3.1.4. - Classification of Words in Subjects

In order to classify words into categories, MeSH classification from the National Library of Medicine (NLM) was used as a guideline. All single word MeSH headers (or their synonyms as defined by the NLM annotators) were selected and then the stem of the word was computed using Tree-Tagger. The words present in the corpus of 104 articles were ordered by frequency and all words occurring more than 200 times were selected. Those matching the selected single-word MeSH headers from six categories (A, B, C, D, E, and G; Anatomy, Organisms, Diseases, Chemicals and Drugs, Techniques and Equipment, and Biological sciences respectively) were selected as belonging to those classes. In order to avoid possible misannotations, words matching more than one category were discarded. Manual analysis of the resulting table of associations was carried out in order to check the associations and to make new ones. A new class not present in MeSH (the X class of "Units, Dimensions, & Parts") was generated in order to include a large number of terms mainly present in the Methods section.

3.2. - Definitions of precision, recall and F-measure

The precision and the recall of can be defined as follows.

Recall = TP/(TP+FN) and Precision = TP/(TP+FP)

Where, TP, TN, FP and FN are true positives, true negatives, false positives and false negatives.

F-measure = $\frac{(1+\beta^2)\times \mathrm{Prec}\times \mathrm{Rec}}{\beta^2 \times \mathrm{Prec} + \mathrm{Rec}}$,

The parameter β can be used for weighing precision and recall terms, but we give equal weight to precision and recall and hence take β=1. Thus, the F-measure is the harmonic mean of precision and recall.

3.3. - Predicate argument structure analysis for written texts in molecular biology

3.3.1. - Selection of Verbs for PAS analysis

The research in molecular biology is multi-faceted and new concepts are added in the literature continuously. However, the areas of cellular signalling, gene expression, regulation and disruption of gene expression, are very important for the larger community of investigators involved in basic biomedical research and those performing high-throughput experiments. These events are discussed throughout the different parts of papers as possible cause of development and disease states of different organisms. While most of the vocabulary found in research articles is similar to that of general English, some verbs are domain-specific in nature. Verbs that are used for describing molecular events in biology were the focus of the analysis. Hence, verbs involved in the above-mentioned processes (events) and present in the literature with high-frequency were chosen.

3.3.2. - Selection of Example Sentences for PAS analysis

Majority of the published IE systems use MEDLINE abstracts as a source for text. The use of abstracts is advantageous because they have highest density of keywords compare to other section of research articles. On the other hand, mining of biomedical texts should scale-up to analyze full-text articles where the most detailed descriptions are contained along with supporting evidences, comparisons to published work in the same area of research and background information, etc. (Mizuta and Collier, 2004) Introduction and Discussion sections apart from the abstracts of articles may be viewed as interesting source of important biological information (Shah et al., 2003). Thus, PAS was defined by analyzing sentences from MEDLINE and from all article sections except the Method section from papers in journals like Proceeding of National Academy of Science , Nucleic Acid Research, EMBO journal and Journal of Virology. Analysis of sentences from various sources would help achieve usage agreement and good PAS frames. Sentences from the Method section are not used in this analysis as they are limited in terms of biomedical information, have generic written styles and verb sense usage tend to overlap with general language.

3.3.3. - Use of parsers reduces manual work

At least 10 sentences per predicate were selected to determine the PAS frame of the verb under study. A sentence under investigation was parsed using Connexor Parser (Tapanainen and Jarvinen, 1997) that uses Functional dependency grammar (FDG), to give parse tree, word, lemma, syntactic function and dependency links between words. Such information helped in determining the existence and boundary of each argument present in a sentence. The parse tree served as a useful guide helping in the manual analysis, but was not considered a gold standard (Figure 3.31). However, usage of the parser considerably reduced the manual work involved in defining arguments.

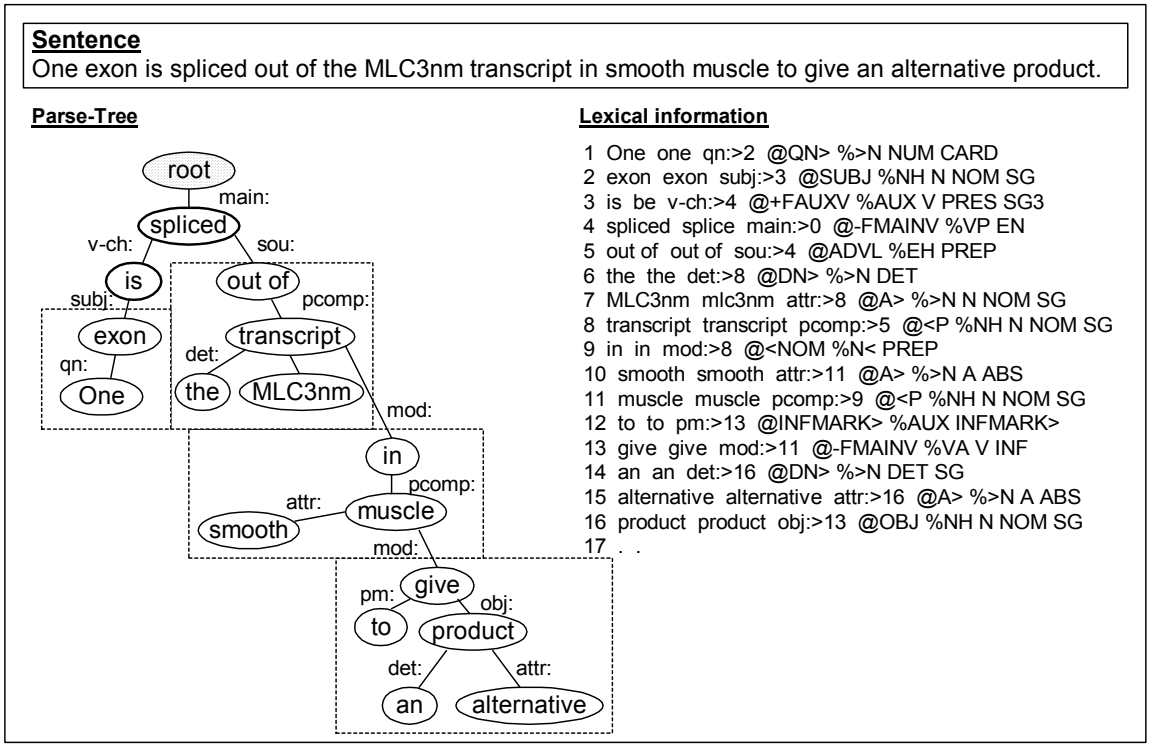

Figure 3.31 - The parse tree generated by the FDG parser: The parse tree generated using an FDG parser provides information at different level. It provides words, lemma, part of speech tags (noun, verbs etc.), syntactic information and relations (subject, object etc.) and also word dependency information.

3.4. - Semi-automated generation of the database of transcript diversity

3.4.1. - Description of transcript diversity in abstracts

In the eukaryotes multiple transcripts are generated and expressed from a single gene by mechanisms including alternative splicing (Graveley, 2001; Graveley, 2002), alternative polyadenylation (Edwalds-Gilbert et al., 1997), and differential promoter usage (Landry et al., 2003). Sentences describing generation of TD may contain descriptions of transcript diversity generating mechanisms, experimental methods, species, physiological conditions, isoform specificity etc (Figure 3.41; categories 1-3).

Category 1

1. The $[\text{NAALADase/PSMA1}]_1$ gene is known to produce multiple $[\text{mRNA splice forms}]_2$ ($[\text{PSMA1a and PSMA1b}]_1$).

2. It had been assumed that the [human]$_3$ [brain and prostate]$_4$ express transcript variants of [p73]$_1$ generated by [differential promoter usage]$_2$.
3. One of these genes encodes [two]$_6$ different enzyme forms, alpha and beta, due to [the differential usage of first exons]$_2$.
4. The newly defined region of [Hu-K4]$_1$ contains [an intron that may be alternatively spliced and seven polyadenylation signal sequences]$_2$.
5. [EDNRDelta3]$_1$ generates the same amino acid sequence at the C terminus, but utilizes [the polyadenylation signal within the open reading frame]$_2$, resulting [in a shorter 3' UTR]$_5$.
6. The larger clone has 5' and 3' ends that are identical to the smaller clone but also has [an alternatively spliced 1.9-kilobase internal segment]$_2$.

Category 2

7. A [HPFK-M]$_1$ cDNA clone lacking [the sequences corresponding to exon IX]$_5$ was isolated from [human]$_3$ [fibroblast]$_4$ (IMR-90) library, suggesting that [HPFK-M]$_1$ transcript may be [alternatively spliced]$_2$.
8. Soluble [Fc gamma receptors]$_1$ are produced by [cleavage of the membrane receptors or by alternative splicing]$_2$.
9. [Northern hybridization analysis and RT-PCR]$_7$ suggests that the soluble and membrane bound forms of [human]$_3$ [AmP]$_1$ are products of [two distinct genes or, through alternative splicing]$_3$, have different [C-terminal sequences]$_5$.

Category 3

10. In this study, we have identified [three]$_6$ [Skn-1]$_1$ isoforms, which encode [peptides with various N termini]$_5$.
11. These [two]$_6$ [hRPB3]$_1$ mRNA species differed in [3' UTR region length]$_5$, the longer transcript containing the AU-rich sequence motif that mediates mRNA degradation.
12. If this question is correct, the observed differences in [amino acid sequences]$_5$ of [protein phosphates 2]$_1$ could be explained by the existence of different mRNAs for gamma and gamma' chains.
13. [Northern blot analysis]$_7$ detected 2.4 kb and 3.2 kb mRNA transcripts of [Ccd1]$_1$ in all tissues examined.
14. [Gene expression analysis using cph genomic fragments from normal and neoplastic cells]$_7$ identifies a number of transcripts including a major mRNA of 2.5 kb as well as several smaller transcripts.
15. There were [tissue-specific]$_8$ differences in the size of [MAP4]$_1$ mRNA transcripts in [human]$_3$ [brain]$_4$ tissues as well.
16. All [six]$_6$ mRNAs of [Pot-1]$_1$ like gene were present in the samples analyzed.

Category 4

17. A G to T mutation in exon 6 results in an in-frame termination codon in eight Hispanic patients from Colorado and New Mexico.
18. Northern analysis and RT-PCR detected aberrant splicing and mutations of TEG101 in human breast cancer cell lines.
19. We report on molecular cloning of a novel human cDNA by its interaction with the splice factor SRp30c in a yeast two-hybrid screen.
20. All exon-intron boundaries agree with GT-AG rule.
21. Using RT-PCR analysis, we show that human 20alpha-HSD, and PGFS mRNAs express ubiquitously, while DD4 mRNA is restricted to the liver.
22. Regions of strong divergence between chicken fast C-protein and human slow C-protein may represent differences in C-protein isoforms.
23. Identification of I-plastin, a human fimbrin isoform that is expressed in intestine and kidney.

Figure 3.41 - Example sentences from MEDLINE describing transcript diversity: Example sentences from the training set, describing generation of transcript diversity (categories 1-3) and negative sentences (category 4). Alternative transcripts are generated by many mechanisms or combinations of them. Hence, the SVM classifier has to learn multiple patterns apart from their syntactic variants. The sentences are

classified in to various categories and semantic patterns are hand-labeled from 1-8. Please see table 4.31 for the pattern labels.

Aberrant transcripts are generated and expressed in disease conditions. Their description may be found in literature describing clinical studies, involving cell lines and tissue samples from patients. Also, the sentences describing expression of a single gene or its products or mechanism of splicing are common in literature. The sentences describing negative conclusions about AS, protein isoforms that may be generated by different gene paralogs, and expression of aberrant transcripts, were taken as negative training examples (Figure 3.41; category 4). These sentences show similar word usage to sentence in catergoris 1-3, making the learning task more challenging.

3.4.2. - Definition of Sentence classification task for inductive learning

The sentence classification task was carried out using inductive machine learning on a training set with labelled examples. During inductive learning the learner ***Ł*** is given a training set ***S*** containing ***n*** examples $(\vec{x}_1, y_1), .. , (\vec{x}_n, y_n),\ y \in \{-1,+1\}$, drawn independently and identically distributed according to an unknown but fixed distribution. Each example consists of a text feature vector $\vec{x}$ and its class label y. The learning task involved the maximization of correct class labels. In such a set up, the classification performance of any method depends upon the quality of features presented to the method and the various learning parameters. Thus, the procedure of feature extraction and parameter estimation is very important for machine learning. The learning performance of a trained classifier is assessed on a set of examples it hasn't seen before. The learning process is repeated until the classifier achieves a satisfactory performance.

3.4.3. - Training corpus and pre-processing for sentence classification

Generation of transcript isoforms diversity is a part of gene expression process. Hence, abstracts with MeSH terms describing gene expression where retrieved and sentences describing TD were chosen for creating the training corpus. A total of 4240 sentences describing generation of multiple alternative transcripts in natural states and 13,520 negative sentences were taken from article titles and abstracts. Aberrant transcripts are generated due to phenomenon including mutations, and nucleotide inversions that lead to diseases. Such sentences are considered negative sentences for the task as the focus was to extract natural transcript diversity.

The Oak system (http://nlp.cs.nyu.edu/oak/; Sekine S., unpublished) was used to split abstract text into sentences. Sentences were tagged with Tree-tagger to assign part of speech tags to the words (Schmid, 1994). Sentences were broken down into words and stemmed to act as primary features to learn from. Stop words, words with certain part of speech tags, and words occurring with very low frequencies (< 5) were removed from the list of words composing the input feature set. The resultant of pre-processing is a file contained all the words (bag of words) occurring in the corpus with a total of 23,742 features. A

'vocabulary' file was generated by manually inspection and removal of non-essential words from the first file to result in 9590 features for this set.

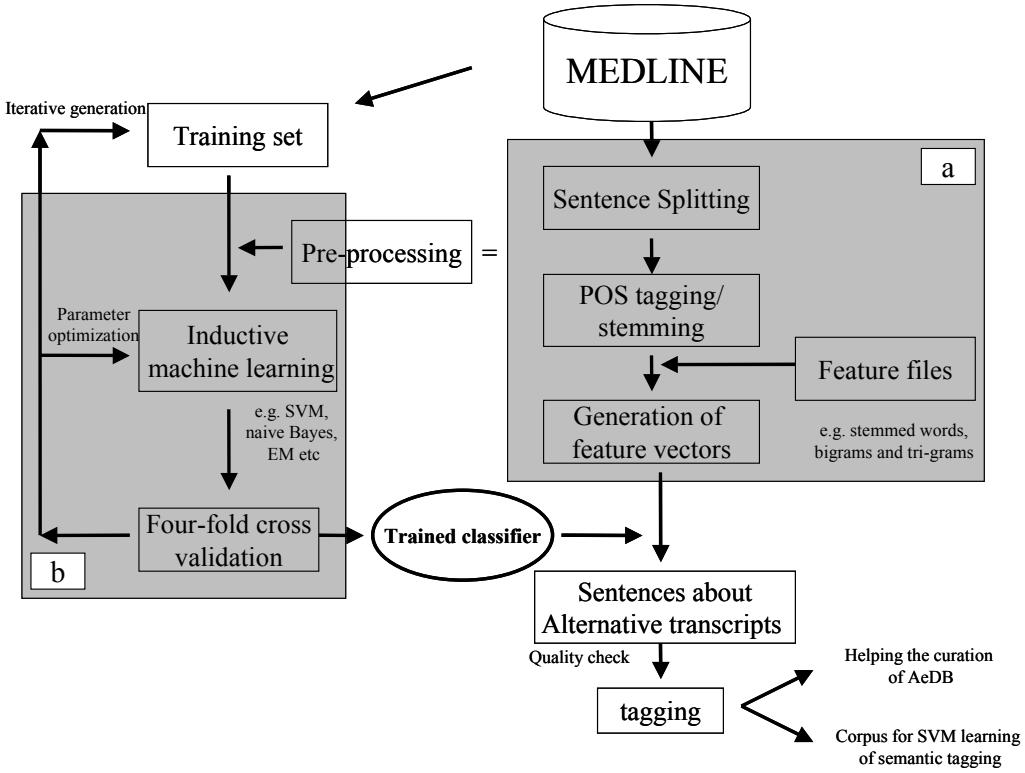

Figure 3.42 - Flowchart of the sentence classification procedure: The procedure has two main modules and other accessory modules. The module marked with *a,* is a pre-processing module. This module is used to convert text into feature vectors that can be used for inductive learning or for classification once the classifier is trained. The module marked with *b*, is the learning module. Inductive learning accompanied iterative generation of training set and parameter optimization in order to get a good performance.

Classification experiments with naïve Bayes, EM, Maximum Entropy, KNN and tf*idf and its variants were performed with the Bow toolkit (http://www-2.cs.cmu.edu/~mccallum/bow/). SVM implementations from the package SVM^{light} was used (http://svmlight.joachims.org). Please see (Joachims, 2001; Mitchell, 1997) and (Ribeiro-Neto, 1999) for a detailed discussion of the methods used here. Also, see Appendix for a short introduction to SVM and of other machine learning algorithms. The training procedure is summarized in Figure 3.42.

3.4.4. - Set for benchmarking of recall for SVM classifier

The classifier performance of the classifier trained to extract only natural transcript diversity was benchmarked against the MeSH annotations in MEDLINE. MEDLINE 2004 contained 8133 abstracts with the MeSH term 'alternative splicing' assigned to them. But only a subset of abstracts provided information about physiologically relevant (natural) transcript diversity. For example, 1725 of these abstracts also contained the MeSH term 'mutation', usually referring to cases of aberrant transcripts and 489 abstracts were without any text. Hence in order to maintain consistency, we removed 2214 abstracts from the list and used remaining abstracts while benchmarking for the recall.

3.4.5. - Mapping of sentence classification results to Sequence databases

The sentences extracted by the SVM classifier were mapped to gene entries in databases like SwissProt, RefSeq, and GenBank using literature entries in those databases. This mapping was done carefully to avoid one sentence/abstract mapping to multiple genes, a situation that may arise in case when literature in a single abstract is attributed to multiple gene entries in the database. The mapping procedure allows entries in LSAT to be linked with sequence databases and thus an access to knowledge stored in those databases.

The success in assigning gene, species, and event mechanisms is as follows (Figure 3.43). 46% of all abstracts were directly mapped to literature entries in the sequence databases like Swissprot (Boeckmann et al., 2003), RefSeq (Pruitt and Maglott, 2001) and GenBank (Benson et al., 2004). A further 15% of all abstracts were assigned the database identifiers using a gene tagger (Mika and Rost, 2004) and the species name extracted from the sentences and/or from the MeSH terms. This mapping was carried out with a synonymous list. However, only 54% of all abstracts could unambiguously be assigned to unique specie (category a in the bottom; Figure 3.43). Rest of the abstracts may have gene and species information but they could not be assigned to a sequence database. Using this mapping procedure 674, 637, and 359 annotations were added for AS for human, mouse and rat genomes, respectively.

3.4.6. - Quantifying the gain in gene annotation

To quantify the gain in gene annotation, first, the sequence information was mapped to the Medline identifiers from the SVM classification using literature entries in Swissprot, Refseq and GenBank. Second, sequence containing entries for human, mouse, and rat genes present in our results and in those databases were mapped to Ensembl (Birney et al., 2004) gene identifiers using the EnsMart system. Then these annotations were compared to that of Swissprot and RefSeq to identify genes that missed the manual curation of AS. Misannotations that arise due to single literature entry mapping to multiple database entries were rejected. Hence, these annotations are highly significant.

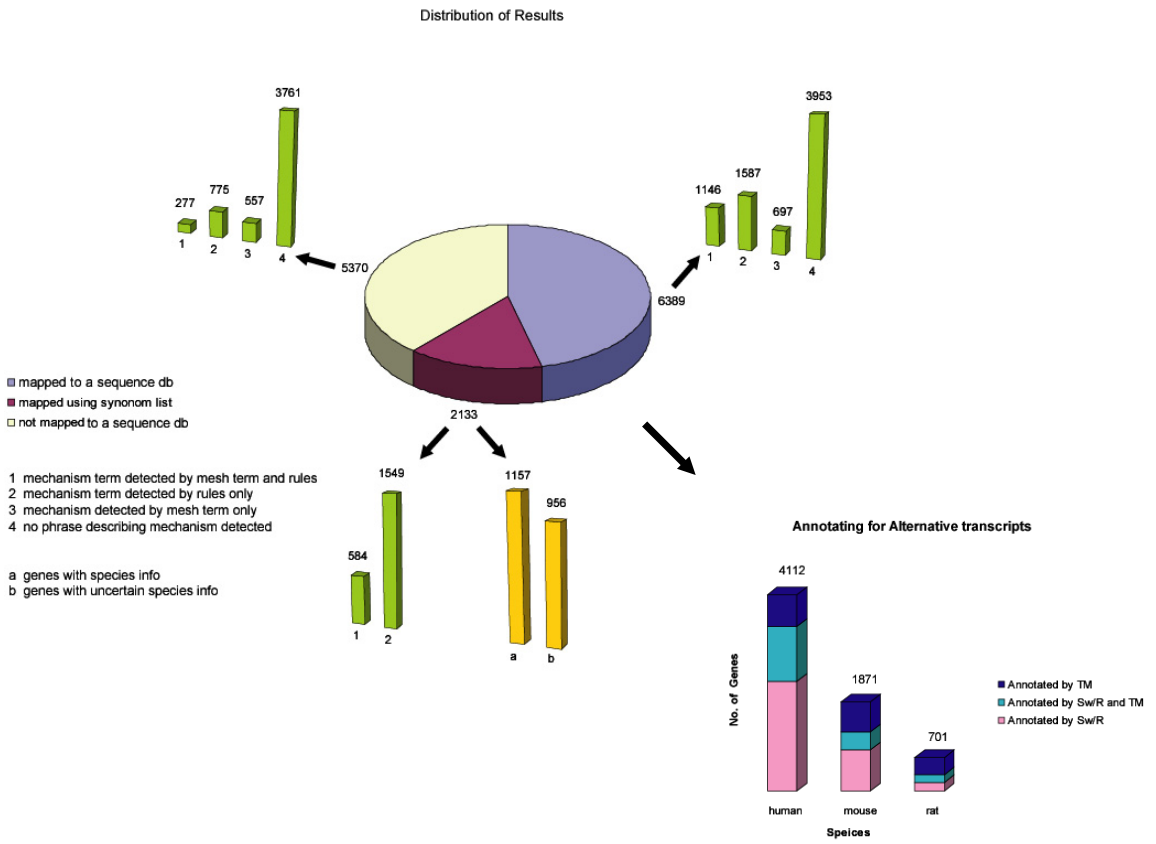

Figure 3.43 - Distribution of results: The pie chart in the middle shows the number of abstracts that could be mapped to sequence databases using literature entries and synonymous list and those that couldn't (clockwise). The bar graphs with categories 1-4 shows number of abstracts in which mechanism could be assigned to genes extracted from those abstracts. MeSH terms and species information was used to identify gene studied in the abstract (bar graph with categories a, and b). Using literature entries present in Swissprot, RefSeq and GenBank databases extraction results were mapped to to Ensembl genes for human, mouse and rat genomes. Annotation increase obtained is shown in the bar graph.

3.4.7. - Merging multiple syntactic patterns to define semantic categories

For example, in the sentence, 'Northern blot analysis detected the presence of a 2.4kb transcript and a 3.2 kb transcript in brain, liver and pancreas', the phrases 'Northern blot analysis' and 'brain, liver and pancreas' would serve the role of arguments to the verb *detect* with semantic labels of *experimental methods* and *tissues*, respectively. It is clear that variation of the sentence as 'Detection of 2.4 kb and 3.2 kb transcripts present in brain, liver and pancreas by northern blot analysis' would not change the semantic role assigned to constituent 'northern analysis' and 'brain, liver and pancreas'. At the same time in sentence, 'Using RT-PCR and nucleotide sequencing, alternative splicing was confirmed in liver, brain and testis', phrases 'RT-PCR and nucleotide sequencing' and 'liver, brain and testis' would serve roles of *experimental methods* and *tissues*, respectively.

3.4.8. - Rules for extracting semantic categories

For example, a rule to find out the role of the variable region in alternatively spliced transcripts in terms of structure or function could be summarized as follows: "*Take Noun phrase chunks right to different forms of verbs 'lack' (Figure 3.41; sentence 4) and 'differ'. Terminate when any of the end condition is encountered"*. The end condition includes encounter of end of line, break in the sentence, different forms of 'be', words like 'through', 'due to' and 'because'. The rule for extracting experimental methods can be described as follows: "*Take chunks left to the different verbs 'show' and 'detect' (Figure 3.41; sentence 4, 6, 8, and 9) containing certain keys words (e.g., PCR or blot). Take the chunks to the right if passive form of verbs is used*".

Apart from the phrases extracted using predicate argument structure analysis, event mechanisms were extracted based on bi-gram and tri-gram phrase lists. Tissue specificity was identified by tagging the word 'specific*' that may follow the tagged tissue name or part of the word describing the tissue (e.g. brain-specific). Similarly, 'number of isoforms' was extracted by the fact that such numbers always preceded the tagged event mechanisms. Tissues were tagged using a dictionary compiled from Swissprot and Refseq. Gene names were tagged using an entity tagger (Mika and Rost, 2004).

3.4.9. - Benchmarking of the tagging performance

From the sentences retrieved by the SVM classifier, instances of eight semantic categories were extracted with rules (see above). Performance (precision and recall) of this tagging rules were evaluated by manually inspecting 300 randomly selected sentences for each category (see Table 5).

3.4.10. - Associating TD-generating mechanisms with organ systems.

The significance of the association of each TD-generating mechanism with each organ system was evaluated using the Hypergeometric distribution. The p-values were corrected for multiple testing by calculating the false discovery rate using the Benjamini-Hochberg formula (Reiner et al., 2003). Total 14 significant associations (out of 45) were found at a 5% false discovery rate, three of which were also significant at a 1% false discovery rate.

IV. Results

4.1. - Analysis of full text articles with keywords

Most applications of information extraction from the biomedical texts use the Abstract of the publication. Abstracts are good for this purpose because they synthesize the content of the article and they are available in public databases. However, nowadays most journals are available in electronic version, and thus full text articles can be used for information extraction.

It is obvious that the full text of an article contains more information than its Abstract. However, in approaching full text analysis several problems must be tackled. The storage of full text articles requires more disk space and the analysis needs more computational capacity. An Abstract, as a summary, contains a high frequency of relevant terms and relationships, but this may not be the case of the rest of the article. Other questions regard the quality of the information carried by different sections of an article. First of all, is the information in full text organized enough so that it can be extracted? Secondly, different kind of information (for example, gene and protein names, tissue names, organisms, experimental conditions, etc.) may be located in different parts of the article. Or it could be that a word has a different meaning depending on the section where it is located (the word has a context dependent meaning). For example, regarding gene names, those found in the Methods section may refer mostly to analytical tools rather than being relevant to the biological phenomenology described in the whole article. In summary, it would be good to quantify and qualify the information in a full text article before embarking in large scale extraction of particular items of information.

The work was aimed to compare the information carried by different sections of a paper. To simplify matters, the work focused on the extraction of relevant words (keywords) regarding objects as they represent a logical view of a given document. A scoring scheme was devised to identify keywords and it is described in Methods.

4.1.1. - Performance at keyword detection

The performance of the keyword detection was evaluated by comparing selected keywords to the MeSH (Medical Subject Headings) terms attached to the articles by annotators at the National Library of Medicine. There are 18.6 MeSH terms per article on average. MeSH terms composed of only a single word (6.80 terms on average) were selected for comparison, as keywords are at the level of single words. It was noted that the most unspecific (for example, *animal*) were often not present in the text and thus could not be matched by a keyword as opposed to species names (*mouse, mycobacterium, human*), or anatomical terms (*hippocampus, cerebellum, breast*). Of those single-word MeSH terms, 4.91 were found on average in the article (as nouns), and 2.22 were among the set of selected keywords (above $K>=0.3$). Obviously, a more accurate comparison to MeSH terms would require the detection of word phrases (bigrams, and

trigrams), but this is out the scope of this work. The recall when matching the original MeSH terms went down from 6.80 / 470.6 = 0.72 in the dictionary of 470.6 different nouns present in an article to 6.80 / 66.6 = 0.33 in the 66.6 keywords selected. However, since the size of the list of all nouns found in an article is much larger than the number of keywords, the precision in matching the MeSH terms of an article increased from 4.91 / 470.6 = 0.010 to 2.22 / 66.6 = 0.033.

4.1.2. - Keyword selection by section

The number of words selected upon a threshold in the K value varies for different sections (Figure 4.11). The first observation was that there were a small number of words that have much better K scores than the rest. This means that the organization of words makes it possible to extract keywords for all five sections considered.

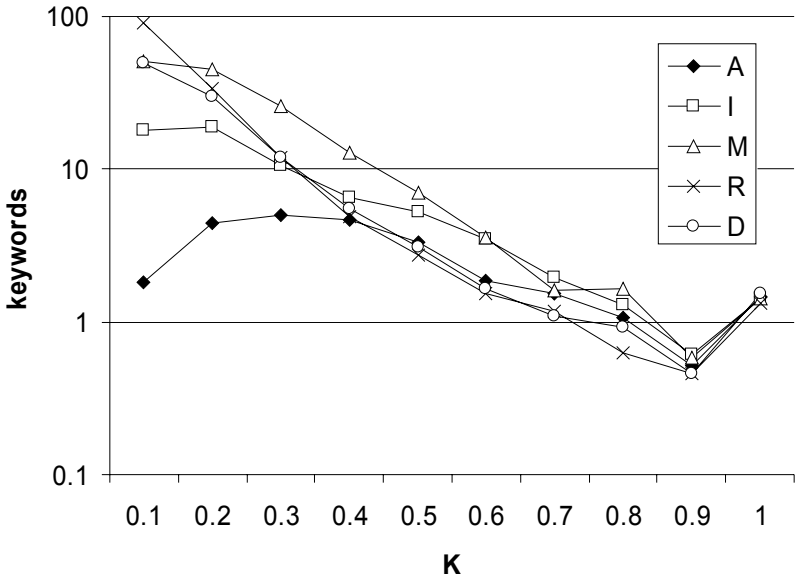

Figure 4.11 - Distribution of keywords by article sections: Average number of keywords versus the threshold *K* for A, I, M, R, and D sections. The average number of nouns per section is, A = 52, I = 171, M = 404, R = 600, D = 331

The number of words selected was very similar for all sections for very high values of *K* (above 0.8). Above a threshold on *K* (*K* >= 0.5; see Table 4.11) the resulting number of keywords were quite similar for Introduction and Methods (around 15 for each) with the other three sections producing around nine keywords. However, if one accounts for the size of the sections it is obvious that the frequency of keywords (selected with *K*>=0.5) per noun was the best in the Abstract (0.18), followed by the Introduction 0.08), with Methods, Results, and Discussion lagging behind. This justifies data mining strategies that

focus in the analysis of Abstracts in order to minimize computational work. However, this result already indicates that not all keywords are in the Abstract, and that therefore mining the rest of the article may be useful.

	All	K>=0.3	K>=0.4	K>=0.5
A	52.17	19.44	14.42	9.77
I	171.32	31.03	20.47	14.00
M	404.19	54.24	28.50	15.80
R	599.98	24.74	12.74	7.85
D	331.04	26.16	14.25	8.75

Table 4.11- Keyword selection per section: Average number of nouns per section (all), or number of those selected as keywords for three different thresholds on the *K* score

4.1.3. - Sections display heterogeneous information

As a way to see how heterogeneous the information is between different sections, the keywords that were common between sections were examined. The results indicated that, typically, not many keywords were common between sections and those present were not very relevant. Even for a low threshold of *K*>=0.3, there is on average only one such general keywords per article. Those are often non-informative words such as "gene", or "protein". This indicates that the information is unevenly distributed across the sections of the article, that is, different sections contain different kind of information.

The heterogeneity of the information by section with the keywords selected (for K>=0.5) for a particular article is illustrated here (Figure 4.12). This work deals with an exon loss resultant of a mutation in the *Nf1* gene of mouse that produces learning deficits (Costa et al., 2001). The only keyword present in every section is the organism under study, the *mouse*. If the Methods section is excluded, only one single more keyword (*mutation*) is selected. Other three-section overlaps give more interesting keywords such as the name of the gene under study (*Nf1*, *neurofibromin*), a domain contained in the resulting protein (*GAP*), the method for testing the learning performance of mice (*maze*), or the resulting phenotype (*impairment, lethality*). Keywords unique to different sections tend to correspond to the different information contained in each section. For example, the keywords unique to the Methods section deal with reagents and techniques (*antibody, amersham, tris, primer*).

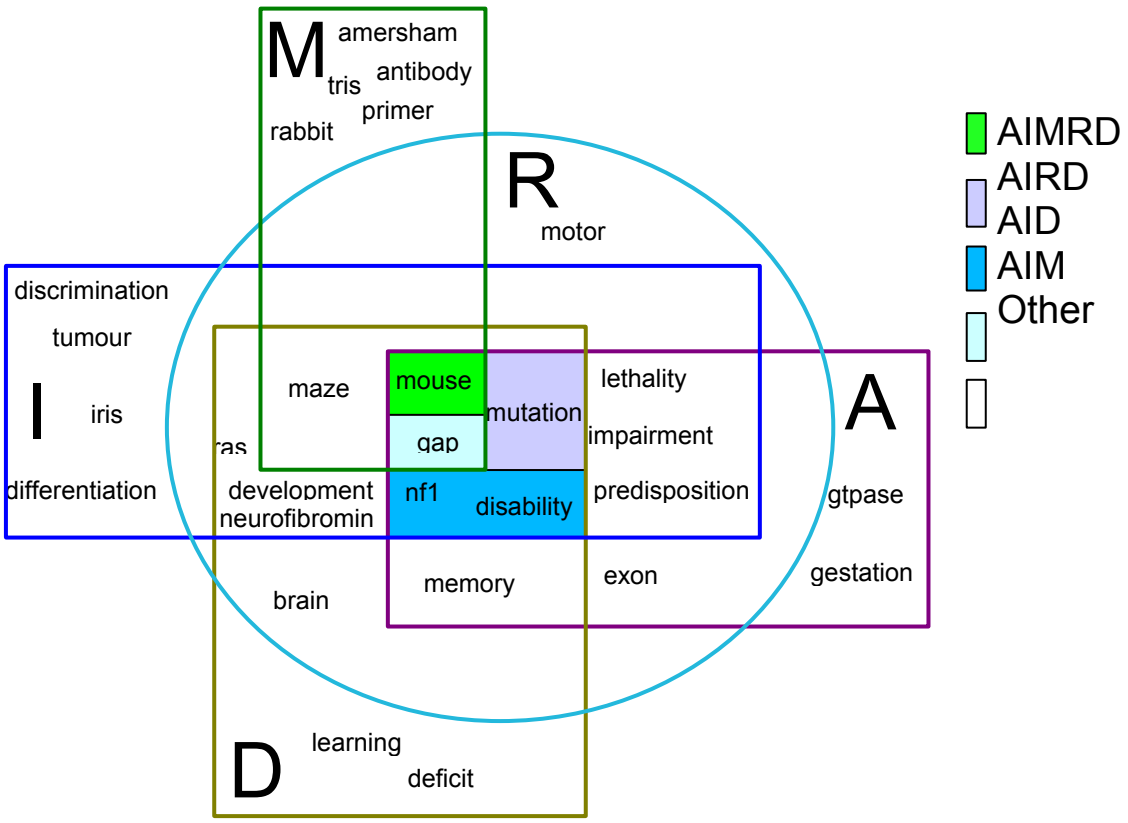

Figure 4.12 - Example of keywords selected for one article: The keywords selected for an article (Costa et al., 2001) with a *K*>= 0.5 are represented as they appear in the different sections of the article

In order to quantify the differences and similarities of content across the article the number of keywords that are shared between different sections were compared (Table 4.12). The values indicate that the Methods section is the most different of all. In Methods, the content is usually focused on the techniques and protocols used, and not so much on the biological phenomena that is the main subject of the article. This alone explains why those keywords present in every section (for example protein, gene) are scarce and uninteresting.

Regarding similarities between sections, A, I, and D are evenly similar among them, and R is the closest to M, as shown when plotting the distance matrix of Table 4.12 as a dendogram (Figure 4.13). The Results section is expected to be closest with the Methods as there the protocols used become prevalent just like the Methods. The Discussion focuses again on the biological results (stressing their relation to the current knowledge) without detailing the techniques that have already been explained in Methods and justified in Results. This result indicates that each section contains certain keywords that are unique to the section. In the following pages differences in content between sections are characterized.

	A	I	M	R	D
A		2.01	0.92	1.77	2.20
I	2.01		0.81	1.34	2.02
M	0.92	0.81		1.55	1.02
R	1.77	1.34	1.55		1.99
D	2.20	2.02	1.02	1.99	

Table 4.12 - Average number of keywords (*K*>=0.5) shared by two sections.

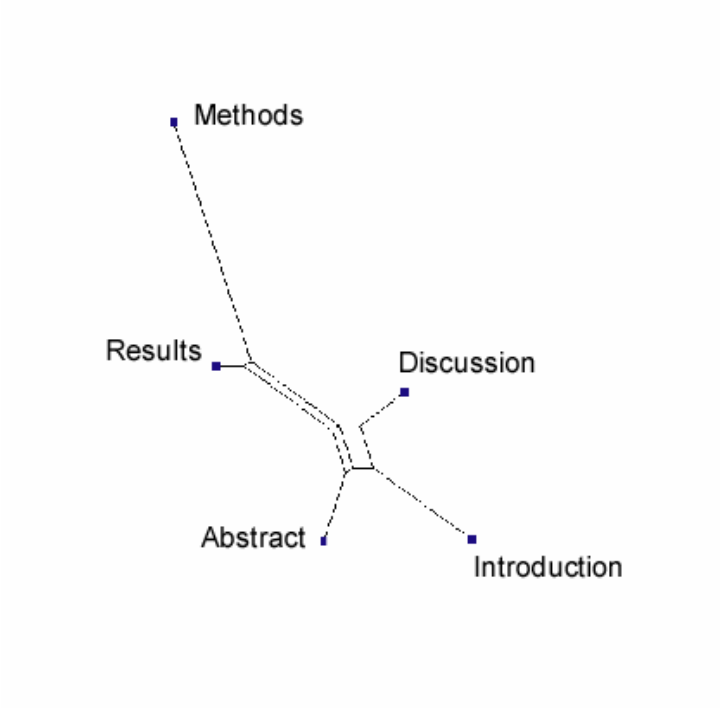

Figure 4.13 - Comparison between article sections: Similarity and differences between standard sections of full-text articles regarding the keyword contents.

4.1.4. - Qualitative analysis of subjects per section

A set of words present in the corpus of 104 articles (not necessarily selected as keywords) were classified in seven categories to analyze further the kind of information present in each of the sections. In order to do so as unambiguously as possible, the words (nouns) that matched MeSH descriptors consisting of a single word and belonging to only one major MeSH category, were used (see METHODS). An

additional category not present in MeSH, that of "Units, Dimensions, & Parts" was defined in order to account for many terms that are currently not MeSH terms but are of interest.

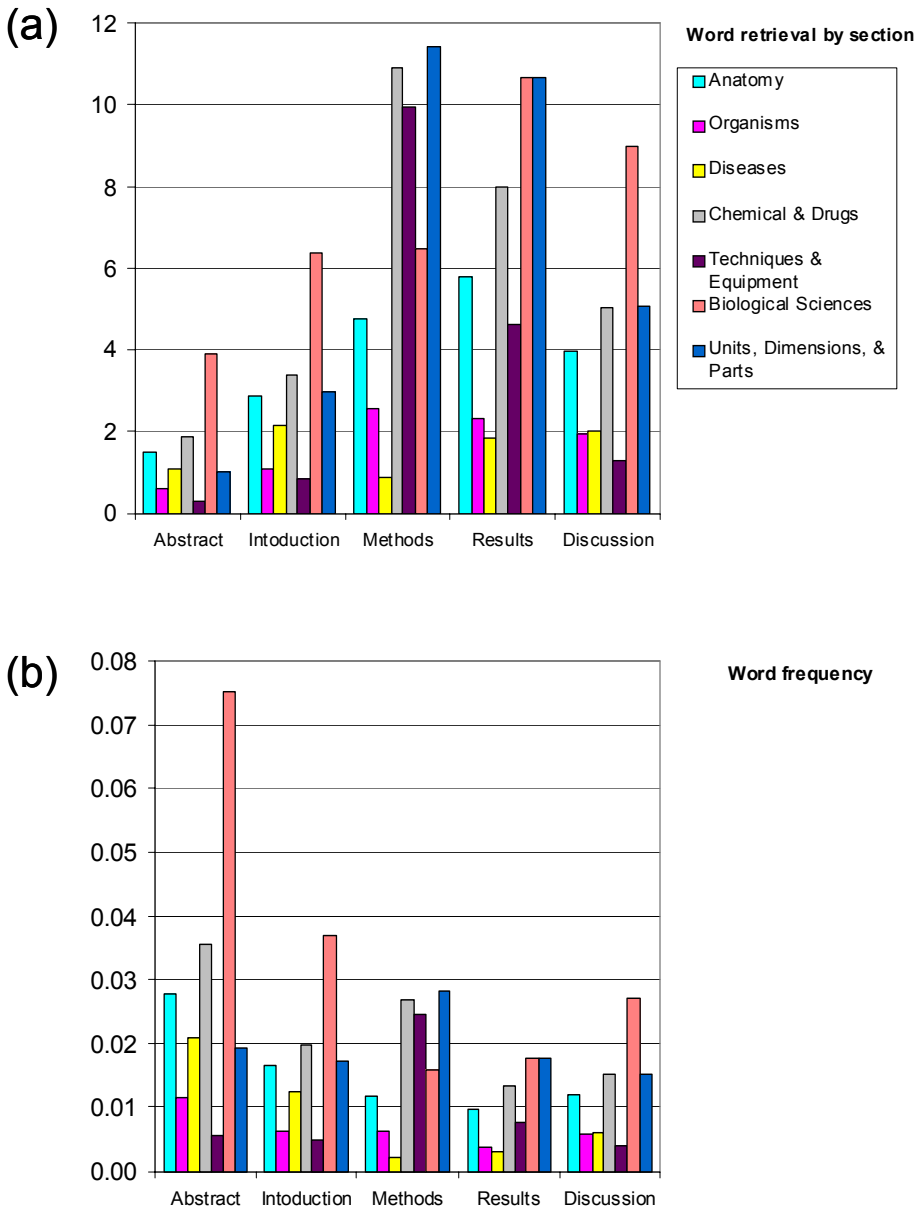

Figure 4.14 - Word categories present in the five sections under analysis: Classes according to MeSH are A (Anatomy), B (Organisms), C (Diseases), D (Chemicals & Drugs), E (Techniques & Equipment), G (Biological Sciences). An additional class X was defined in this work (Units, Dimensions, & Parts). The number of words used for the analysis was 36 (class A), 14 (B), 11 (C), 47 (D), 33 (E), 41 (G), 49 (X). (a) Average number of occurrence of words of each subset per section. (b) Frequency of words of each subset per total number of words for each section.

The results (Figure 4.14a) indicate that the large sections are a good source of keywords, expectedly the Methods with many terms related to techniques. Introduction, Results and Discussion contain a good deal of information regarding diseases. However, again, the Abstract section is shown as the best source for most subjects regarding frequency of keywords (Figure 4.14b) except for those typical of the Methods section (Techniques & Equipment; Chemicals & Drugs).

4.1.5. – Analysis of distribution of gene names

Detection of gene and protein names (NE extraction) is a very important subject, broadly used for the detection of macromolecular interactions (see Introduction), and one objective of this work concerned with the relevance of matching gene names in different sections of an article to study the context information, the distribution of gene names across sections was examined.

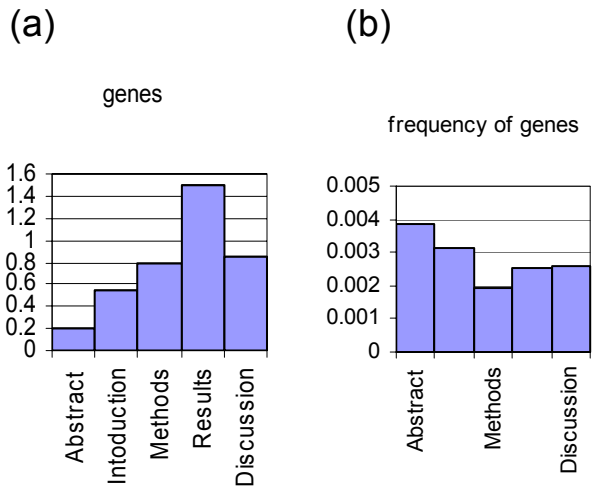

Figure 4.15 - Distribution of gene names across sections: (a) Average number of different gene names per section from the set of 224 genes. (b) Frequency of different gene names per total of nouns for each section

As mentioned in Introduction, gene name identification is not an easy task and frequently these names tend to be ambiguous. For example, there are gene names called *Not* or *That*. Shorter names (e.g. *A6*) can also be a problem. In order to avoid the ambiguity in the analysis, a set of 539 genes whose names composed of three letters followed by one single digit was selected from the Swiss-Prot database (Bairoch and Apweiler, 2000). A total of 224 gene names out of the 539 were matched in 76 of the 104 articles. The Results section was the one with more gene names (Figure 4.15a). Again, the Abstract, and then the Introduction, were the sections with the highest frequency of these names (Figure 4.15b).

The context of gene names that were exclusively mentioned in the Methods section was checked manually in order to study the problems that affect gene-name identification if context is ignored (even

when using gene names apparently easy to recognize). Of the 224 genes, just 24 were mentioned in the Methods section of the corresponding 14 articles and not elsewhere.

In five of the 14 articles, the name was referring to a non-gene object (three restriction endonucleases, a vector name, and a fibroblast cell strain). In four articles, the gene was mentioned in a technical context (usually, the gene mRNA level was used for analysis of cell state) and no biological process involving the gene was described. In only four articles the mention of the gene name was relevant. Additionally, it was noted that of these 24 gene names, at least two (*Pbp2*, *Pom1*) could refer to two non-homologous (unrelated) genes, and another one (*Sac1*) to four; such synonymous gene names make gene identification difficult. In summary, one should be careful with the context in which gene names appear. Extreme caution should be applied with gene names appearing uniquely in the Methods section.

4.2. – PASBio: Towards event extraction from biomedical texts

The complexity in IE increases as one moves from NE extraction to relationships and event extraction. There are already several reports in the literature for describing methods for NE identification from biomedical texts. It has been a shared task for at least two community-wide efforts (see Introduction). There have been ongoing efforts for relationships extraction since the beginning of NLP in biology. Event extraction, which is considered the highest level IE, is the task on which this part of the thesis concentrates.

Event extraction involves the filling of an event template that makes use of the results from entity recognition. As mentioned in the introduction, traditional IE methods utilizing rules based only on syntax suffer from performance degradation due to the fact that a single event can be written in a variety of syntactic forms. Moreover, relationship extraction from complex sentences in technical and scientific texts requires deeper knowledge of sentence semantics. Regular expression based methods that use shallow parsing based argument chunking capture only weak notions of argument structures. On the other hand, PAS frames formalize the notion of arguments centered on a predicate. Thus, semantic extraction templates based on PAS would perform better at the task of event extraction. Utilization of PAS frames is the ideal solution to the problem of performance degradation facing the IE methods caused by the existence of multiple syntactic patterns (see Introduction). Therefore, sentences from biomedical corpus were analyzed for semantic roles and argument sets of interesting predicates in biomedical text and the resultant PAS frames were compared to those proposed for English from the newswire text representing the usage in general domain.

The resultant PAS frames are available in the form of a database of predicate argument structures in biology (PASBio) at http://research.nii.ac.jp/~collier/projects/PASBio (Wattarujeekrit et al., 2004).

4.2.1. - Mapping from surface structures to PAS

Mapping from the surface structures to PAS could be illustrated with the example mentioned earlier, (a) "*Peter sprayed water on his flowers*." and (b) "*Peter sprayed his flowers with water*." Both sentences can be mapped into the PAS of the verb *spray*, which indicates the event of "applying thin liquid to surface" with 3 required arguments (agent, liquid, and surface). In both the sentences the constituent "*Peter*" plays the semantic role of an *agent* who does the action, "*water*" that of the *liquid* used in the event, and "*his flowers*" is conceived as the *surface* getting wet. It should be noted that the position of "water" in sentence (a) is that of a direct object following a verb which belongs to a part of a prepositional phrase as in sentence (b). Similarly, a sentence from biological corpus such as "*One exon is spliced out of the MLC3nm transcript in smooth muscle to give an alternative product.*" could be conceptualised into PAS relationship as follows (Figure 4.21).

A syntactic parser would assign the sentence constituents "*One exon*", "*is spliced out*", "*of the MLC3nm transcript*", "*in smooth muscle*", and "*to give alternative product*" with their syntactic categories as *noun phrase*, *verb*, *prepositional phrase*, *prepositional phrase*, and *verb phrase* respectively. At the syntactic relations level, "*One exon*" is the *surface subject* of the passive form verb "*is spliced out*" and other constituents play the role of *complements* (Figure 4.21).

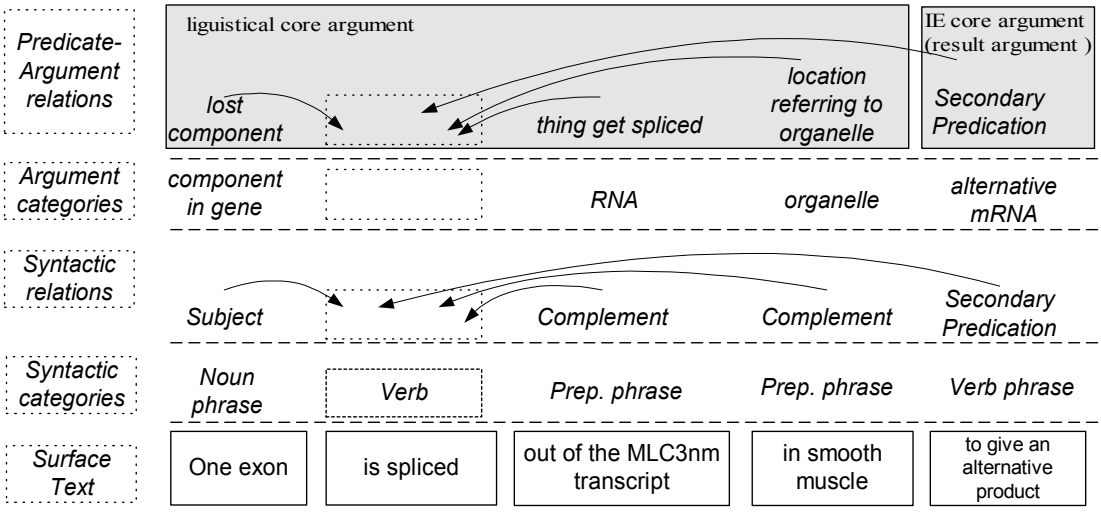

Figure 4.21 - Syntactic and semantic level description of the surface text: There are different levels of understanding the surface text. Syntactic categories provide a syntactical class for each constituent of the sentence. Syntactic relations describe the syntactical function of each constituent of the sentence to the predicate of the sentence. Argument categories offer the semantic concept for each constituent of the sentence. Predicate-argument relations level helps in understanding the semantic role played by each constituent or argument related to its predicate

Above the syntactic levels, there are semantic levels including the argument categories and predicate-argument relations. At the argument categories level "*One exon*", "*the MLC3nm transcript*", "*smooth muscle*" and "*alternative product*" constituents belong to the domain concept classes of *a gene product (mRNA)*, *tissue* and *alternative mRNA* respectively. The predicate-argument relation level contains the most abstract information. Semantic roles played by other constituents to the verb indicating the event are represented at this level. Thus, the process of *removal of an exon from mRNA* is indicated by the verb *splice out*. Here, the verb arguments play the semantic roles of *lost component* ("One exon"), *entity getting spliced* ("the MLC3nm transcript"), *location referring to tissue* ("smooth muscle"), and *secondary predication - showing purpose or reason in this example* ("to give an alternative product") and suggest alternative splicing of MLC3nm mRNA. The argument "*to give an alternative product*" is assigned the semantic role *secondary predication* because this argument by itself is capable of instantiating a PAS frame and therefore it is considered to be a core argument.

The semantics of a sentence relate in complex ways to the syntax of the sentence as illustrated by various semantic and syntactic levels (Figure 4.21). Using this layered approach different surface forms describing a given event can be mapped into the same PAS. Thus, PAS could be helpful for IE to overcome the syntactic variation problem in molecular biology.

4.2.2. - Defining predicate-argument structures for molecular biology

In molecular biology, genes and their products are at the centre of the study, as a set of these molecular entities dictate, and their products carry out, different functions at the cellular level and the combined effects can be seen at the organism level. Hence, in the sentences in biomedical literature genes or its products are described as agents participating in some events, with the help of appropriate verbs indicating specific events (Figure 4.22). Different molecular-level or phenotypic effects are described as the other arguments of such events. As described above, PAS is a representation of semantic relationships between arguments with specified roles and a verb relating to a particular event narrated in a sentence. Thus, PAS should be a natural choice for IE especially event extraction in molecular biology.

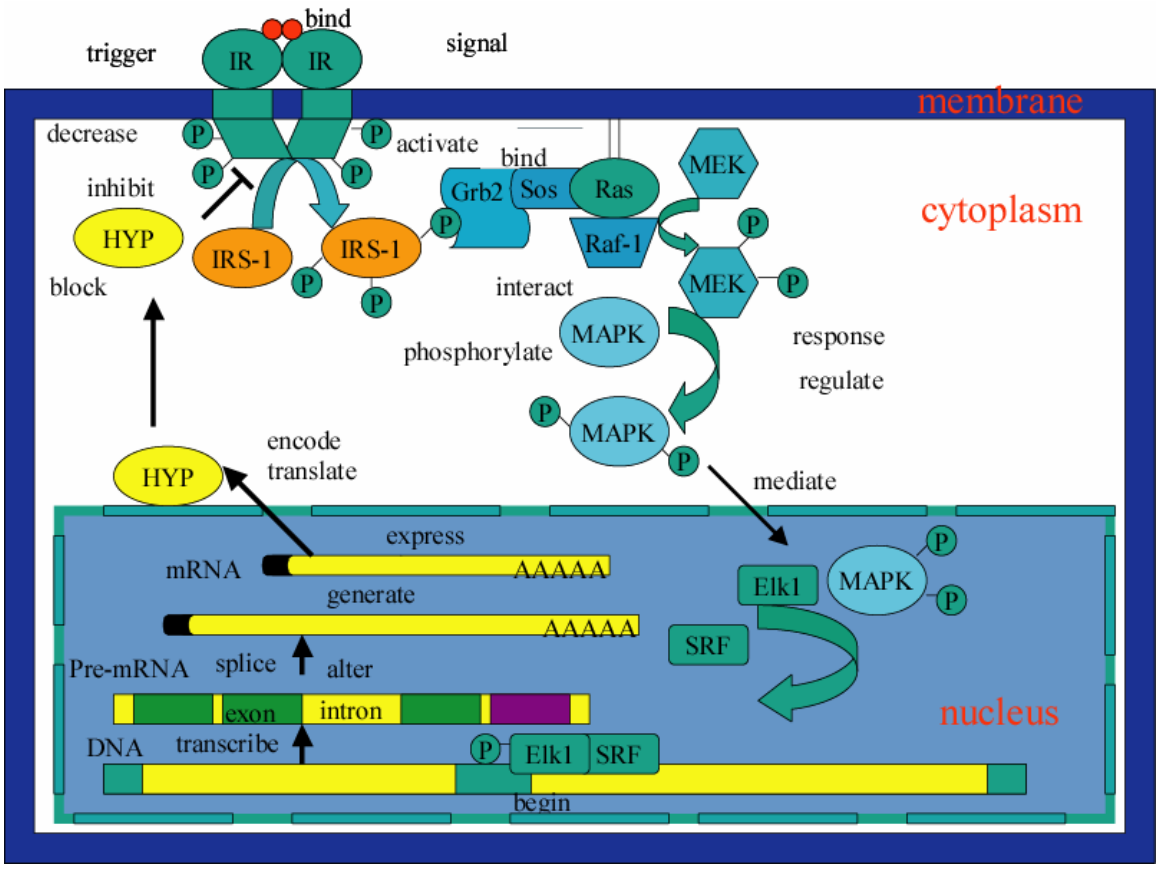

Figure 4.22 - Molecular events as described by associated predicates: A hypothetical signal transduction pathway of an idealized cell is shown here. The signal is triggered at the outer membrane by the ligand binding to a receptor dimer. This signal is mediated (by various proteins) to the nucleus of the cell through various events (protein-protein interactions, phosphorylation etc.) and leads to initiation of transcription of a gene. Following transcription, splicing and translation, the protein product inhibits receptor signalling. Thus, it regulates its own expression via a negative feedback loop. The direction of information flow is shown with arrows. Cell compartments, molecular entities and predicates describing various events are shown. The predicates analyzed in this work aim to cover events in gene expression, regulation and signalling processes

4.2.3. - Guidelines for defining PAS

For defining PAS for biomedical texts, PropBank's scheme was adapted with necessary changes. To define PAS for any verb, a survey of the usage of the verb and presence of various arguments was made from a set of sample sentences in a representative corpus. A verb may have several senses depending upon the usage (e.g., express- to speak or send quickly). In PASBio, these senses were divided with the aim of

obtaining fine-grained semantic senses using the WordNet lexicon (Miller, 1990). Each PAS frame in of PASBio contains a set of core arguments and auxiliary arguments. An argument is declared as a core argument if it is important to complete the meaning of the event described in the sentence. Nevertheless, if an argument is important but there is no evidence that the argument exists together with the predicate in at least 20% of the example sentences, it may not be considered a core argument. *Arg X* (with *X*, a cardinal number, starting from *0* and incremented with each additional argument) and *ArgR* are used with a mnemonic label for any core arguments. The difference between *Arg X* and *ArgR* is illustrated for the PAS of *mutate* in the next section (Figure 4.23). The mnemonic label is a short description of the semantic role played by the argument. Biological function and usage of the argument were considered to describe semantic roles in PAS. No attempt was made to ensure the consistency of mapping between argument labels (argument name) and the roles (the mnemonic labels) played by the arguments, except *Arg0*. *Arg0* was reserved for the argument playing the semantic role of an *agent*. In cases where *Arg0* was not present, the first core argument was labelled *Arg1* in the PAS frames of such verbs. See PAS frames for *mutate* (Figure 4.23), *express* (Figure 4.27) and *transform.02* (Figure 4.28) as examples.

The sentence constituents identified as adverbial, negation and modality were annotated with the tag *ArgM-type* (*ADV* and *MAN* in the case of adverbial, *NEG* in case of negation, and *MOD* in case of modality), in addition the core arguments. However, only adverbials in terms of adverbs were considered to be annotated as *ArgM-MAN* (for a manner adverb) or *ArgM-ADV* (for other types of adverbs). If adverbial phrases or adverbial clauses are mandatory for expressing events indicated by particular predicates, they are defined as core arguments within the PAS frames. For example, an adverbial phrase playing the role of a locative modifier was included in the set of core arguments for the predicate *initiate*. (Refer to example sentence "Apparently HeLa cells either initiate transcription *at multiple sites within RPS14 exon 1*."). Moreover, adverbs that play roles of manner modifiers (e.g. *normally*, *genetically*, etc.) were distinguished from other adverbs.

A manner adverb deserves special distinction from other adverb types because it shows how a certain action is performed. Annotating a manner adverb is very important to understand facts in a sentence from biomedical texts. For example, "*normally*" in the sentence "Mice have previously been shown to develop *normally*" is necessary for IE in order to understand that the development process was normal. Other types of adverbs, for example aspectual modifiers that give information about whether the temporal information about affairs (e.g. "*still*" in the sentence "The mice were *still* developing normally even after the deletion Msp1 gene."), adverbs acting as frequency modifiers (e.g. "always" in the sentence "We found IL-2 expression *always* on the plasma membrane."), adverbs acting as focusing modifiers like *even, only, also,* and *too* (e.g. "The transcription is initiated *only* in female blastoderm embryos."), were tagged as *ArgM-ADV*. In case of negation and modality, *ArgM-NEG* and *ArgM-MOD* are given directly to a negator word (i.e. not or n't) and a modal verb (i.e. will, may, can, shall, must, might, should, could and would) respectively. Negations (operating at the sentence level) and modality (operating at various levels) were not defined as core arguments because linguistically neither of them can be considered an argument within the

PAS frames. They are only worth annotating from an IE perspective if they exist in a same clause where a focused predicate exists. Similarly, adverbials were also considered worthy of being annotated when present as they can significantly alter or even reverse the meaning of the sentence.

4.2.4. Examples of defined PAS

Three important cases were examined to assess domain specific behaviour of biomedical predicates with the assumption that domain specific usage of verbs in biology would influence its PAS for biological domain. They are (1) verbs that are rarely used in general language (e.g. *splice*) or have a unique biological interpretation (e.g. *express*, *translate*, etc.), (2) verbs that have a similar meaning used in the general and biological texts but show different patterns of usage (e.g. *alter*, *initiate*, etc.), and (3) verbs that are used with the same meaning and usage style in both domains (e.g. *abolish*, *delete*, etc.). PAS frames proposed by Propbank were taken as a representative for the verb usage in general English. Therefore, PAS frames in PASBio were compared to those in PropBank for a verb under consideration. The results of the comparison fall in to four groups of verbs. They are discussed below:

Verbs with same semantic sense but require more core arguments

As an example, consider the event of mutation, beneficial mutations get selected in the population paving the way for evolution and harmful mutations cause diseases that may be inherited. The verb *mutate* describe the physical changes at the molecular sequence level. PropBank defines two arguments for this verb which are *Arg0: agent* and *Arg1: mutated entity*, but four arguments are needed for a complete PAS frame of the verb *mutate* in biomedical text (Figure 4.23). As mentioned before, *Arg0* is reserved only for the argument playing the semantic role of agent. However, the sentences commonly used to describe *mutate* events are normally written in passive forms. Therefore, the agent of the event is not stated explicitly in most of the sentences.

Hence, core arguments for *mutate* start from *Arg1* as the position for the agent is left empty unless there is mention of a mutagenic agent. *Arg2* describes the NE participating in the event and is analogous to PropBank's *Arg1*. Thus, *Arg1*, *Arg3*, and *ArgR* defined in PASBio for *mutate* are extra arguments compared to PropBank. Arguments *Arg1*and *Arg3* are captured conforming to the linguistic criteria that the semantic role of an element is implied when the element is omitted and that element should be an argument (Meyers et al., 1994). From the biological perspective, these two extra arguments are implied. Noticeably, consequences of the event driven by the verb *mutate* are often seen in examples. Apart from "changes at molecular level" assigned as *Arg3*, the consequence, "changes at phenotypic level" is suggested as *ArgR* (explained below). Sentence 1.1, 1.2, and 1.3 support this explanation.

<table>
<tr><th colspan="2">1) Predicate: MUTATE</th></tr>
<tr><th>Argument Structure for Biology</th><th>PropBank Argument Structure</th></tr>
<tr><td>Arg1: physical location of mutation //exon,intron, domain//
Arg2: mutated entity // gene //
Arg3: changes at molecular level
ArgR: changes at phenotype level</td><td>Sense = to undergo and cause to undergo mutation
Arg0: agent
Arg1: entity undergoing mutation</td></tr>
<tr><td colspan="2">Match to MUTATE senses in WordNet: sense 1 – undergo mutation</td></tr>
<tr><td colspan="2"><u>Sentence 1.1</u> The exon 5 mutated allele with the premature translation termination resulted in severe deficiency of Hex A.

Pred: mutate
Arg1: exon 5
Arg2: allele
Arg3: [with] the premature translation termination
ArgR: resulted in severe deficiency of Hex A

<u>Sentence 1.2</u> The gene mutated in variant late-infantile neuronal ceroid lipofuscinosis (CLN6) and in nclf mutant mice encodes a novel predicted transmembrane protein.

Pred: mutate
Arg1: -
Arg2: gene
Arg3: [in] variant late-infantile neuronal ceroid lipofuscinosis (CLN6) and in nclf mutant mice
ArgR: encodes a novel predicted transmembrane protein

<u>Sentence 1.3</u> Transient expression of the exon 8 mutated alpha-chain cDNA in COS-1 cells resulted in deficiency of enzymatic activity.

Pred: mutate
Arg1: exon 8
Arg2: alpha-chain cDNA in COS-1 cells
Arg3: -
ArgR: resulted in deficiency of enzymatic activity</td></tr>
</table>

Figure 4.23 - PAS for mutate, a verb in group A: The PAS of mutate contains more arguments than PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002). Extra arguments responsible for consequences of the event mutate are considered as core arguments as they are often seen in sentences from biomedical documents. WordNet (Miller, 1990) sense 1 – undergo mutation corresponds to the biological sense of mutate. Three sentences in the lower panel illustrate how surface structure was mapped into PASBio's predicate-argument structure.

2) Predicate: INITIATE	
Argument Structure for Biology	**PropBank Argument Structure**
Arg0: agent //gene// Arg1: entity created //transcription or translation// Arg2: specific location on gene //exon or intron// Arg3: location as tissue or cell Arg4: method	Sense = begin Arg0: agent Arg2: theme (-creation) Arg3: instrument
Match to INITIATE senses in WordNet: sense 1 – brought into being	
Sentence 2.1 Apparently HeLa cells either **initiate** transcription at multiple sites within RPS14 exon 1, or capped 5' oligonucleotides are removed from most S14 mRNAs posttranscription. **Pred: initiate** **Arg0: -** **Arg1: transcription** **Arg2: [at] multiple sites within RPS14 exon 1** **Arg3: HeLa cells** **Arg4: -** **Sentence 2.2** I kappa B-epsilon translation **initiates** from an internal ATG codon to give rise to a protein of 45 kDa, which exists as multiple phosphorylated isoforms in resting cells. **Pred: initiate** **Arg0: -** **Arg1: I kappa B-epsilon translation** **Arg2: [from] an internal ATG codon** **Arg3: -** **Arg4: -** **Sentence 2.3** Since RTKs **initiate** signaling by recruiting downstream components to the activated receptor, proteins that are immediately downstreamof an activated RTK can be identified by first identifying sequences in the RTK that are necessary to activate downstream signaling (Schlessinger and Ullrich, 1992; Pawson, 1995). **Pred: initiate** **Arg0: RTKs** **Arg1: signaling** **Arg2: -** **Arg3: -** **Arg4: [by] recruiting downstream components to the activated receptor**	

Figure 4.24 - PAS for initiate, a verb in Group A: The PAS frame of initiate also belongs to group A – same sense, more arguments. Similar to the predicate mutate, additional arguments responsible for spatial information of the event described by intitate are proposed to be core arguments.

The argument *ArgR:results/consequences* is the argument providing information about consequences after the event denoted by the predicate occurs (Figure 4.23). For *mutate*, most of the examined sentences contain *ArgR*, revealing the necessity of it. Moreover, the requirement of this argument coincides with biological observations. Therefore it is considered as a core argument (more precisely an IE core argument) and named as *ArgR* instead of *ArgX* (a core argument from a purely linguistic perspective). This distinction is made under the rule that *ArgX* has to play a role during the event but not after the event. This condition is depicted by a formula like "mutation event = (*ArgX* + mutation + *ArgX*) + *ArgR*". Empirically, *ArgR* is seen with verbs relating to an abnormal biological phenomenon or tissue specific differences. Examples of other verbs that need this argument are *skip*, *delete*, etc.

PAS frame of *initiate* also requires additional core arguments. *Arg2* (Figure 4.24, sentences 2.1 and 2.2) describes the point of transcription initiation and *Arg3* provides information about the tissue/cell where the gene (or product) is expressed. In PropBank, the sentence segment defined by the parser with the LOC (location) tag is considered as non-required element. However, the extraction of spatial information is very important from the biological perspective. Furthermore, a variety of semantic roles are present in the subject position in the sentence from biomedical texts (Figure 4.24). In Sentence 2.1 "*HeLa cells*" is syntactically the subject, which seems to be the agent of an *initiate* event, but at the molecular level only a polymerase bound to a target gene may initiate the transcription. "*HeLa cells*" is annotated as *Arg3* referring to *tissue/ cell (location)* in which the transcription takes place instead of *Arg0: agent*. In sentence 2.2, "*I kappa B-epsilon translation*" is also a subject as in the previous example, but it is "entity created" assigned as *Arg1*. Only in Sentence 2.3 (describing initiation of a signalling event), the subject of the sentence fills the semantic role "agent", so a subject "*RTKs*" can be annotated as *Arg0*. Additionally, the point to note is "the entity created" in sentence 2.3 is different from sentence 2.1 and 2.2 as it is a signalling event that is initiated, but not a transcription or translation event.

Verbs that are used with same semantic sense but take fewer core arguments

The verb *block* both in biomedical texts and in newswire texts has very similar semantics (Figure 4.25). However, an event described by the verb *block* in the biomedical domain may not contain a secondary predication and an instrument most of the time. As mentioned before, a *secondary predication* in itself is capable of instantiating another PAS frame. For example, in the following sentence from PropBank "[*John*$_{\text{Arg0}}$] *blocked* [*Mary* $_{\text{Arg1}}$] *from* [*completing her dissertation* $_{\text{Arg2}}$] *with* [*his constant pestering* $_{\text{Arg3}}$]." the Arg2-secondary predication is annotated for "completing her dissertation" because this contains in itself the PAS of the verb *complete*. In this example, the meaning of the event denoted by *block* cannot be understood completely if the sentence just states as "[*John*$_{\text{Arg0}}$] *blocked* [*Mary* $_{\text{Arg1}}$]." Thus, it is necessary to mention the action being stopped. In contrast, in the biomedical texts, by mention of only the entity being stopped (Sentence 3.1-3.3), an expert reader can understand that the full meaning of the sentence without requiring the presence of a secondary predication. Similarly, an instrument used to block is encoded in the nature of an agent or causer.

In the PAS of *block* (Figure 4.25) core arguments exist only in Sentences 3.1 and 3.2 (the agent is not mentioned). In Sentence 3.3, the word denoting manner (specifically) is marked as a secondary argument using symbol *ArgM-MAN*. The secondary argument is not considered in the PAS as a core argument but it is used when a sentence is annotated. This secondary argument is not important or not expected. The semantic role of this argument is independent of the verb. PropBank also uses this scheme.

<table>
<tr><td colspan="2">3) Predicate: BLOCK</td></tr>
<tr><td>Argument Structure for Biology</td><td>PropBank Argument Structure</td></tr>
<tr><td>Arg0: agent, causer
Arg1: theme //entity being stopped//</td><td>Sense = oppose, halt, stop
Arg0: agent
Arg1: theme (action or object being stopped)
Arg2: secondary predication
Arg3: instrument</td></tr>
<tr><td colspan="2">Match to BLOCK senses in WordNet: sense 3 – stop from happening or developing</td></tr>
<tr><td colspan="2"><u>Sentence 3.1</u> Tagetin is more specific for distinguishing between different RNA polymerases because it blocks RNA polymerase during elongation.

Pred: block
Arg0: it
Arg1: RNA polymerase during elongation

<u>Sentence 3.2</u> Membranes were blocked in TBST (Tris-buffered saline, 0.05% Tween-20) containing 5% bovine serum albumin (for anti-phosphoryrosine blots) or skimmed milk and probed with antibodies.

Pred: block
Arg0: -
Arg1: Membranes

<u>Sentence 3.3</u> Mutations at the 3’ splice site that specifically block step II do not affect the association of hPrps 16 and 17 with the spliceosome, indicating that these factors may function at a stage of step II prior to recognition of the 3’ splice site.

Pred: recognize
Arg0: Mutation at the 3’ splice site
Arg1: step II
ArgM-MAN: specifically</td></tr>
</table>

Figure 4.25 - PAS for block, a verb in group B: The PAS frame for block, belonging to group B – same sense, fewer arguments is proposed here. Though this verb is used with the identical meaning in both biomedical and business news corpus, the set of arguments differ. Also, the use of ArgM-MAN is illustrated here.

Verbs that are used with same semantic sense and have identical frames

Specialization of domain hasn't affected PAS frames of verbs in this group.

5) Predicate: CONFER	
Argument Structure for Biology	**PropBank Argument Structure**
Arg0: agent //mechanism, process, entity// Arg1: given biological property Arg2: entity receiving biological property //gene product, cell//	Sense = grant, give Arg0: agent Arg1: gift Arg2: given to
Match to CONFER senses in WordNet: sense 2 – present	
Sentence 5.1 Besides these side chain interactions with the 06-alkyl group, structure-based analysis of mutational data suggests that substitutions at Gly156 and Lys165 **confer** resistance to 06-BG through backbone distortions. **Pred: confer** **Arg0: substitutions at Gly156 and Lys165** **Arg1: resistance** **Arg2: [to] 06-BG** **Sentence 5.2** The portion of the STATs **conferring** specificity for either a MAPK or a MAPK substrate kinase (MAPKAP) has not been determined. **Pred: confer** **Arg0: The portion of the STATs** **Arg1: specificity** **Arg2: [for] either a MAPK or a MAPK substrate kinase (MAPKAP)**	

Figure 4.26 - PAS for confer, a verb in group C: Predicate confer belong to group C – same sense, same structure, so their structures constructed in PASBio are as same as in PropBank as shown in Frame 5 and Frame 6, respectively

An example of such a predicate is '*confer'* that is used with the meaning "to give to someone or something" (Figure 4.26).

Verbs with domain specific semantics

The word *express* is used in the biomedical texts with the meaning "to manifest the existence of a gene or a gene product" (or detection of the same during experiments) unlike its normal usage with the meaning of "give an opinion or send quickly" (Figure 4.27). The predicate *transform*, is used in biomedical text with two senses: "to cause (a cell) to undergo genetic (or neoplastic) transformation" and "to transfer a gene from source organism into target organism altering the target" (Figure 4.28). Even though the first meaning of *transform* is similar to the sense of "change" found in PropBank, there is still a semantic difference between them. In the biomedical literature sentences describing genetic transformations

(Sentences 8.1-8.3) mention only the agent or causer, the entity getting transformed, and the effects of transformation. The start state of the entity undergoing transformation is not mentioned as it usually refers to a normal condition of the entity. *Transform* in the second sense always occurs in a sentence connected by the preposition *into*, and in the passive voice form, in which no mention is made with regard to the agent.

<table>
<tr><td colspan="2">7) Predicate: EXPRESS</td></tr>
<tr><td>Argument Structure for Biology</td><td>PropBank Argument Structure</td></tr>
<tr><td>Arg1: named entity //gene or gene products//
Arg2: property of the existing name entity
Arg3: location referring to organelle, cell or tissue</td><td>Sense = say (express.01)
Arg0: speak
Arg1: utterance
Arg2: hearer
Sense = send very quickly (express.02)
Arg0: sender
Arg1: thing sent
Arg2: sent to</td></tr>
<tr><td colspan="2">Match to EXPRESS senses in WordNet: sense 5 – manifest the effects of a gene or genetic trait</td></tr>
<tr><td colspan="2"><u>Sentence 7.1</u> Northern blot analysis with mRNA from eight different human tissues demonstrated that the enzyme was expressed exclusively in brain, with two mRNA isoforms of 2.4 and 4.0 kb.

Pred: express
Arg1: the enzyme
Arg2: [with] two mRNA isoforms of 2.4 and 4.0 kb
Arg3: [in] brain

<u>Sentence 7.2</u> Two equally abundant mRNAs for il8ra, 2.0 and 2.4 kilobases in length, are expressed in neutrophils and arise from usage of two alternative polyadenylation signals.

Pred: express
Arg1: mRNAs for il8ra
Arg2: 2.0 and 2.4 kilobases in length
Arg3: [in] neutrophils

<u>Sentence 7.3</u> T cells from double TCR transgenic mice express only one or the other of the two available TCRs at the cell surface.

Pred: express
Arg1: one or the other of the two available TCRs
Arg2: -
Arg3: T cells from double TCR transgenic mice</td></tr>
</table>

Figure 4.27 – PAS of express, a verb in group D: Predicate express is used in biological documents with WordNet sense 5 – manifest the effects of a gene or genetic trait which is totally different from the usage found in business news (i.e. say and send very quickly).

10) Predicate: TRANSFORM.01	
Argument Structure for Biology	**PropBank Argument Structure**
Sense = to cause (a cell) to undergo genetic transformation Arg0: agent/causer of transformation Arg1: entity undergoing transformation Arg2: effect of transformation/end state	Sense = change Arg0: causer of transformation Arg1: thing changing Arg2: end state Arg3: start state
Match to TRANSFORM senses in WordNet: sense 2 – change or alter in form, appearance, or nature	
Sentence 8.1 We and others have found that FGF8b can **transform** the midbrain into a cerebellum fate, whereas FGF8a can promote midbrain development. **Pred: transform** **Arg0: FGF8b** **Arg1: the midbrain** **Arg2: [into] a cerebellum fate** **Sentence 8.2** Phospholiipase D (PLD) is known to stimulate cell cycle progression and to **transform** murine fibroblast cells into tumorigenic forms, although the precise mechanisms are not elucidated. **Pred: transform** **Arg0: Phospholipase D (PLD)** **Arg1: murine fibroblast cells** **Arg2: [into] tumorigenic forms** **Sentence 8.3** Overexpression of the retroviral oncoprotein v-Rel can rapidly **transform** and immortalize a variety of avian cells in culture. **Pred: transform** **Arg0: Overexpression of the retroviral oncoprotein v-Rel** **Arg1: a variety of avian cells in culture** **Arg2: -**	

Figure 4.28 – Two PAS frames for transform, a verb in group D: PASBio has different PAS frames for different senses of transform found in biological corpus. PAS as transform.01 is defined based on the meaning – change or alter in form, appearance, or nature (WordNet sense 2) and transform.02 – change (bacteria cell) into a genetically distinct cell by the introduction of DNA from another cell of the same ore closely related species (WordNet sense 6; next page).

9) Predicate: TRANSFORM.02 = TRANSFORM INTO	
Argument Structure for Biology	**PropBank Argument Structure**
Sense = to transfer gene from source organism into target organism Arg1: entity being inserted Arg2: organism or cell undergoing transformation	Sense = change Arg0: causer of transformation Arg1: thing changing Arg2: end state Arg3: start state
Match to TRANSFORM senses in WordNet: sense 6 - change (a bacterial cell) into a genetically distinct cell by the introduction of DNA from another cell of the same or closely related species)	
Sentence 9.1 This construct was **transformed into** the yeast strain HF7c (Clontech). **Pred: transform into** **Arg1: This construct** **Arg2: the yeast strain HF7c (Clontech)** **Sentence 9.2** For expression of the recombinant protein, pET28a-5 was **transformed into** Escherichia coli strain BL21(DE3). **Pred: transform** **Arg1: pET28a-5** **Arg2: Escherichia coli strain BL21(DE3)** **Sentence 9.3** To generate GST fusion proteins, the relevant DNA fragments were cloned into pGex2T (Pharmacia) and **transformed into** the bacterial strains BL21 or TOPP (Stratagene). **Pred: transform** **Arg1: the relevant DNA fragments** **Arg2: the bacterial strains BL21 or TOPP (Stratagene)**	

4.2.5. - Complexities in Biology Texts

In the discussion so far it has been assumed that the predicate is at the centre of semantic information. However, argument contents can also alter the event description specified by the predicate. This can be illustrated with sentences that describe 'alternative splicing' event. Alternative splicing is used to generate multiple mRNA transcripts from a single gene and hence is a helpful event for increasing the functional complexity of eukaryotic systems.

Consider the following example of a set of sentences that talk about the 'expression' of a single type of mature mRNA generated from 'splicing' of pre-mRNA and generation (and expression) of multiple mature mRNA transcripts with different properties from a single type of pre-mRNA. Sentences annotated follow PASBio's frame for *express*: (a) "*Northern blot analysis with mRNA from eight different human tissues demonstrated that* [*the enzyme* $_{Arg1}$] *was expressed exclusively* [*in brain* $_{Arg3}$], [*with two mRNA isoforms of 2.4 and 4.0 kb* $_{Arg2}$]." and (b) "[*A complementary DNA clone encoding the large subunit of the essential mammalian pre-messenger RNA splicing component 2 snRNP auxiliary factor (U2AF65)* $_{Arg1}$] *has been isolated and expressed* [*in vitro* $_{Arg3}$]*."* Sentence (a) is considered as a sentence denoting the

alternative splicing event but sentence (b) is considered as a negative (not describing alternative splicing) sentence, which talks about expression of an mRNA of a splicing factor.

It would be difficult, based on word contents or regular expression methods, to put these two examples into different 'bins' for alternative splicing events. But the discussion about the length of the two different transcripts in Arg2 (with two mRNA isoforms of 2.4 and 4.0 kb) in the first sentence can be helpful to understand it as a sentence discussing alternative splicing. On the other hand, the later sentence contains all the interesting words (e.g., mRNA, express and splicing) but misses Arg2, hence describes just an expression event.

4.3. - Extraction of information about transcript diversity from MEDLINE

As mentioned in the previous section, generation of alternative transcript diversity is considered a very important task for functional diversity and evolution of eukaryotes (also see Introduction). Alternative transcripts generated by alternative splicing (AS) allow eukaryotes to generate an expanded proteome from a limited gene pool. Differential promoter usage and alternative polyadenylation in synergy with AS may change terminal exons or in general regulate expression of mRNA transcripts (Black, 2000; Edwalds-Gilbert et al., 1997; Zavolan et al., 2003). Therefore, an information extraction tool is much required by the community working on elucidating the extent of usage of these mechanisms and their functional implications in different tissues in a single organism and across species. It will also help computational methods that are under development to map the alternative transcript universe. Thus, it is important to identify descriptions of alternative transcripts from abstracts in MEDLINE. Information including gene names, species, tissues, expression-specificity, event mechanisms, and experimental methods could be extracted to generate a database of events that generate transcript diversity.

4.3.1. - Overall strategy and generation of the database

To extract information about transcript diversity (TD) and associated spatio-temporal information scattered throughout MEDLINE, a composite procedure was devised (Figure 4.31). In the first step, sentences containing TD information were identified within the papers' abstracts (IR/text categorization step). As discussed in Intorduction, rule based methods tend to perform worse than machine learning methods due to existence of multiple syntactic patterns. Moreover, as shown in the section 4.2.5 extracting sentences regarding the transcript diversity would be difficult only on the basis of word contents. Thus, a text classifier based on machine learning methods was trained for the sentence classification task by inductive machine learning (Mitchell, 1997) on an annotated corpus (Joachims, 2001; Nello Cristiani, 2000; Vapnik, 1999). The entire MEDLINE database was processed using the trained classifier in order to identify sentences describing TD within the abstracts.

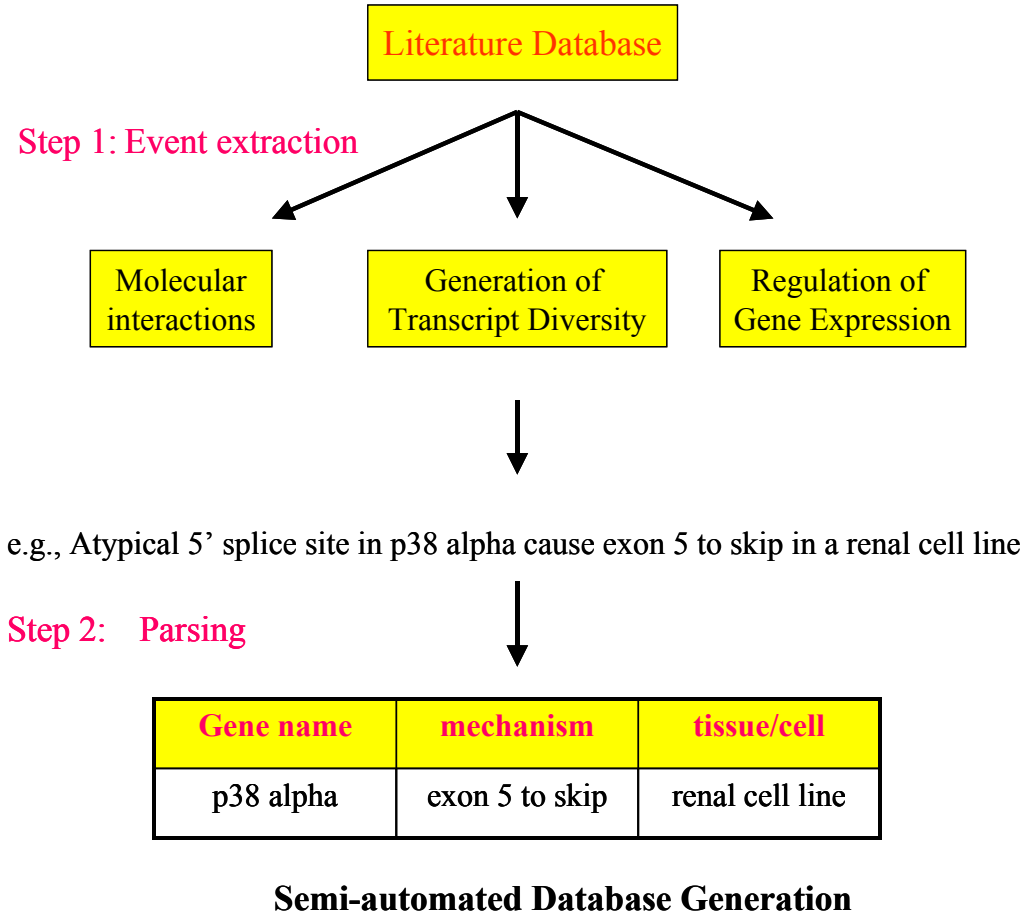

Figure 4.31 - Creating specialized databases for events of interest: A database of physiologically occurring AS events can be generated in two steps. Each step may involve machine learning or rule based methods. The first step involves the identification of sentences from scientific text. These sentences can be parsed in a second step for extracting frequently occurring semantic patterns.

In the second (IE) step (Figure 4.31), sentences were parsed and sentence constituents were assigned different semantic categories (see Methods). Finally, each abstract with information about alternative transcripts (retrieved by the SVM classifier) was mapped to entries in Swissprot (Bairoch and Apweiler, 2000), RefSeq (Pruitt and Maglott, 2001), GenBank (Benson et al., 2004), and Ensembl (Birney et al., 2004) databases, when possible. This not only provides the sequence information at genome, transcript, and protein levels for the genes described in the abstracts but also allow to an interested user to access structural and functional information about these genes stored in various sequence databases. All this information obtained for each PubMed entry constitutes an entry in LSAT (Figure 4.32).

Entry: 60	Pmid 10102990
Title: A novel form of human neuropsin, a brain related serine protease, is generated by alternative splicing and is expressed preferentially in human and adult brain.	
Genbank: AB008390, AB008927	Ensembl: ENSG00000188879.1
Refseq Report:	
Identfiers	NM_144506, NM_144507, NM_144505, NM_007196
species	Human
Gene Definition	Kallikrein 8 (neuropsin/ovasin) (KLK8)
Comment	REVIEWED
Transcript variants	Four transcript variants in refseq
Swissprot Annotations	
Identifier	KLK8_HUMAN
Description	Neuropsin precursor (EC 3.4.21.-) (Kallikrein 8) (Ovasin) (serine protease)
Annotations	Alternative Splicing
Text Extraction Data	
Gene name	neuropsin
Event	Alternative splicing; was a species-specific splice variant
Experimental method	Sequence analysis of the 946 bp genomic DNA spanning the region encoding the insertion sequence
Tissue	brain
Species	Human, mouse
Isoform-number	two
Specificity	Species-specific

Figure 4.32 - An example LSAT entry: A typical database entry is shown in the figure. Apart from the text extraction data, LSAT provides links to PubMed, GenBank, Refseq, Swissprot and Ensembl. The entry for the gene Neuropsin is shown here. This gene is already annotated for alternative splicing in Swissprot. However, text extraction data point to the fact that it is a species-specific splicing absent in mouse.

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 4.31 – Performance of at the extraction of semantic patterns.

Eight different semantic categories describing biologically relevant data were identified in the sentences describing TD among which are event mechanism, species, tissue-specificity, and experimental methods (Table 4.31).

4.3.2. – Experiments on sentence classification

Sentence classification was carried out with the inductive learning procedure to obtain sentences about alternative transcripts. Inductive learning methods learn patterns from the features extracted from the training set and generalize. The generalization performance of many methods degrades when dealing with large amounts of rarely occurring features. Text data is a typical example of this situation sometimes termed as '*the curse of dimensionality'*. Moreover, the process of preparing a reliable training set is expensive and time-consuming. Hence, a good learning method should be able to learn from a small amount of training examples and should be able to handle large number of features. The sentence classification performance of well-known text categorization methods was compared to find the best classification method (Mitchell, 1997). These methods are 1) naïve Bayes, 2) maximum entropy 3) Expectation Maximization (EM), 4) variants of the term frequency-inverse document frequency (tf*idf) methods, 5) K-nearest neighbour (KNN) algorithm and 6) support vector machines (SVM) Using a relatively large corpus of 17,760 sentences, the classification performance of the above mentioned methods was checked with different fractions of the training set in order to choose the best performing method.

Selecting the best sentence classification method

Intuitively, while training with the most basic set, the learning method good at feature selection would outperform the rest. Hence, classification performance of various methods was compared with different fractions of the training set while utilizing the simplest feature set (bag of words; see Methods) that contained more than 23,000 training features.

As expected, the overall classification performance, calculated as the F-measure, improved with the amount of training data (Figure 4.33). However, the SVM and the maximum entropy classifiers consistently outperformed other categorization methods. Also, the SVM classifier with the RBF kernel outperformed that with linear kernel even though text data is hypothesized to be linear. The KNN algorithm with a number of neighbours ranging from 5 to 50 either suffered from memory problems or didn't seem to learn the classification rule. It was clear that the SVM with three different kernels performed better than the other methods and was taken for further characterization.

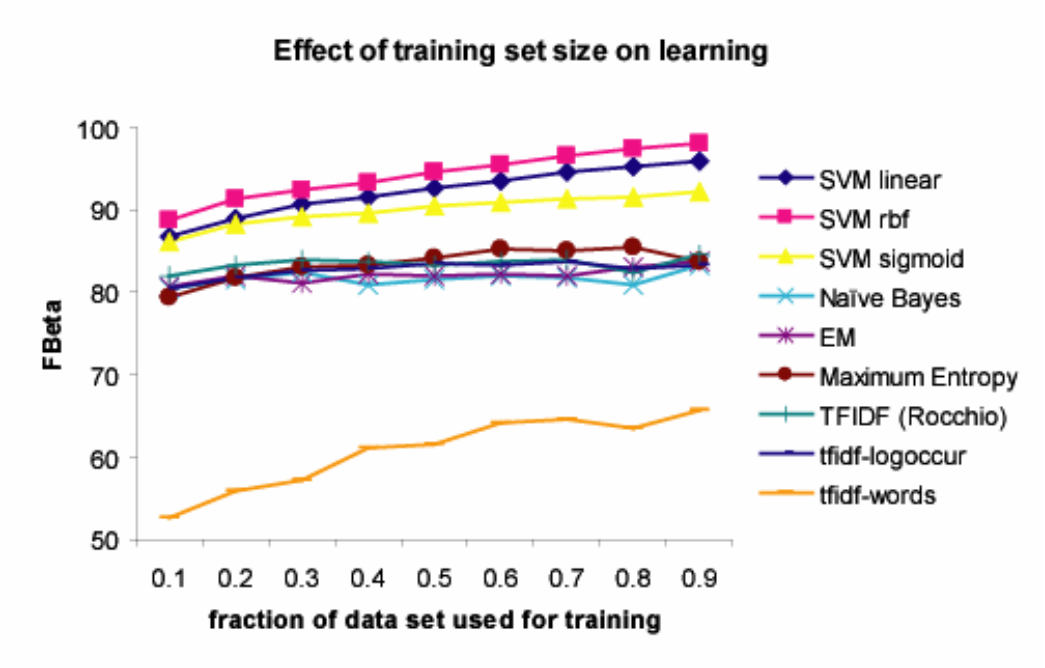

Figure 4.33 - Comparison of various text-categorization methods: The classification performance of various methods (F-measure) was measured with different fractions of training set were used for learning. Each data point here is a mean of results from four cross validations. (EM - Expectation Maximization and tf*idf - term frequency-inverse document frequency methods)

Parameter optimization for the SVM learning

The classification performance of the SVM was explored further with three different kernel functions and associated learning parameters (see Methods). For all kernel functions the value of parameter C in the SVM optimization problem controls the trade-off between the training error and the margin (Joachims, 2001). The value of C depends on the training data and it was determined empirically. Also, the RBF and sigmoid functions have one and two variable parameters, respectively, that can affect the learning process (see Methods). The value for parameter C was characterized with different values of gamma for the RBF kernel and different values of r for the sigmoid kernel with *bag of words* as the input feature set (Figure 4.34). The value of 1.5 for gamma and the value of 10 for C were the best classification parameter values for the SVM with the RBF kernel. Similarly, the value of 0.01 for r and the value of 1000 for C were the best parameter values for the SVM with the sigmoid kernel.

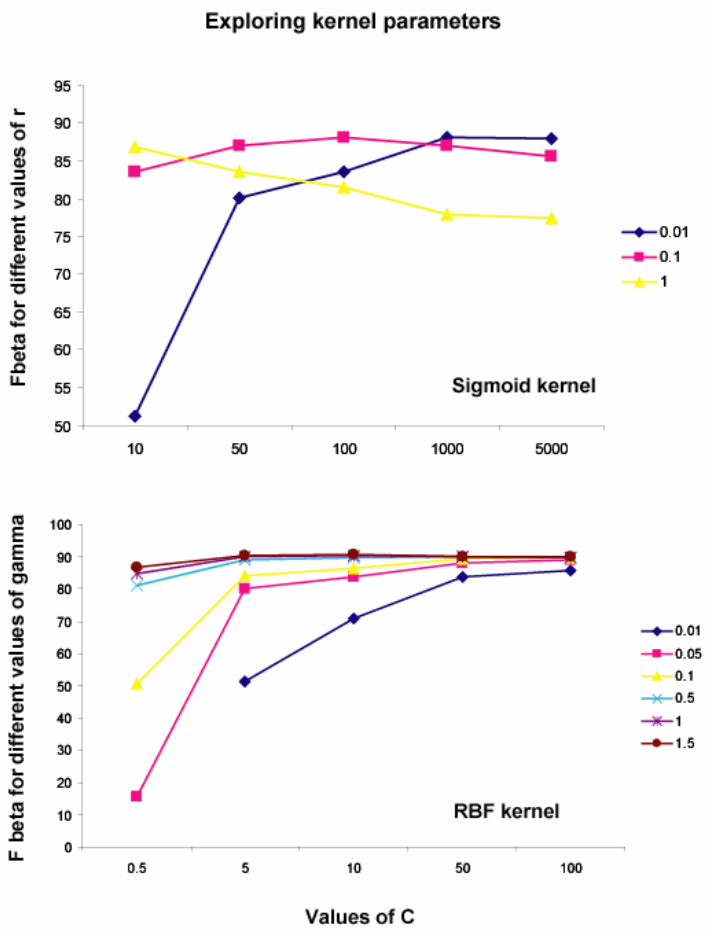

Figure 4.34 - Parameter optimization for SVM learning: Different values of gamma for the RBF kernel and r for the sigmoid kernel at different values of parameter C (figure 3a) were studied as a function of F-measure; 'bag of words' was used as the input feature set. Each data point is an average from four cross-validations.

Feature Enrichment

The process of extracting a rich feature set from the training examples is the most important step in machine learning because methods provided with rich features need fewer training examples and provide better generalization.

Feature enrichment was achieved as follows (Figure 4.35). Many phrases (word bi-grams and tri-grams; e.g., *alternative transcript* or *alternative first exon*; see Appendix) frequently present in the training

sentences were incorporated as additional input features, as a way to add the domain knowledge. Cardinal numbers were summarized as a single feature. In addition, synonyms were defined for the sparsely occurring features (e.g., long transcript, larger transcript and elongated transcript). Two additional feature sets were generated by combining phrases and synonyms with 'bag of words' and 'vocabulary' (see Methods).

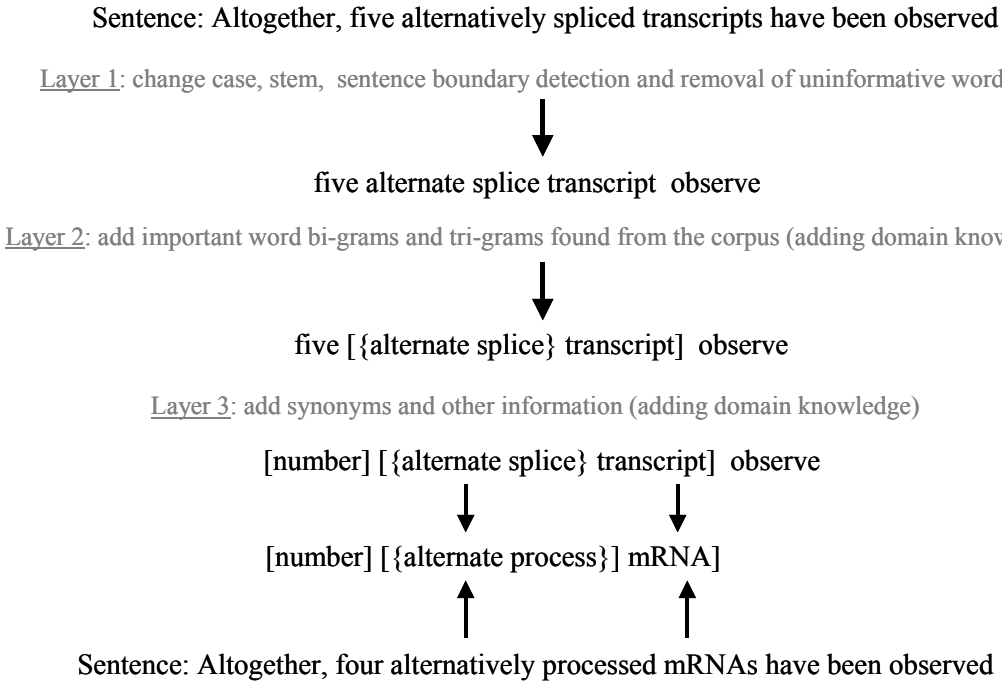

Figure 4.35 – An example of feature enrichment: The procedure of feature enrichment is illustrated here which merges two different sentences to a single patter using three layers of text operations.

The learning performance of the SVM with all three kernels was checked with these feature sets with different values of C (Figure 4.36). The input feature set containing *bag of words and phrases* performed best for SVM with all three kernels. Again, the SVM with the RBF kernel achieved the best performance. It achieved a mean F-measure value of 91% when performing four randomized trials with 60% of total corpus as training set and the rest as the test set.

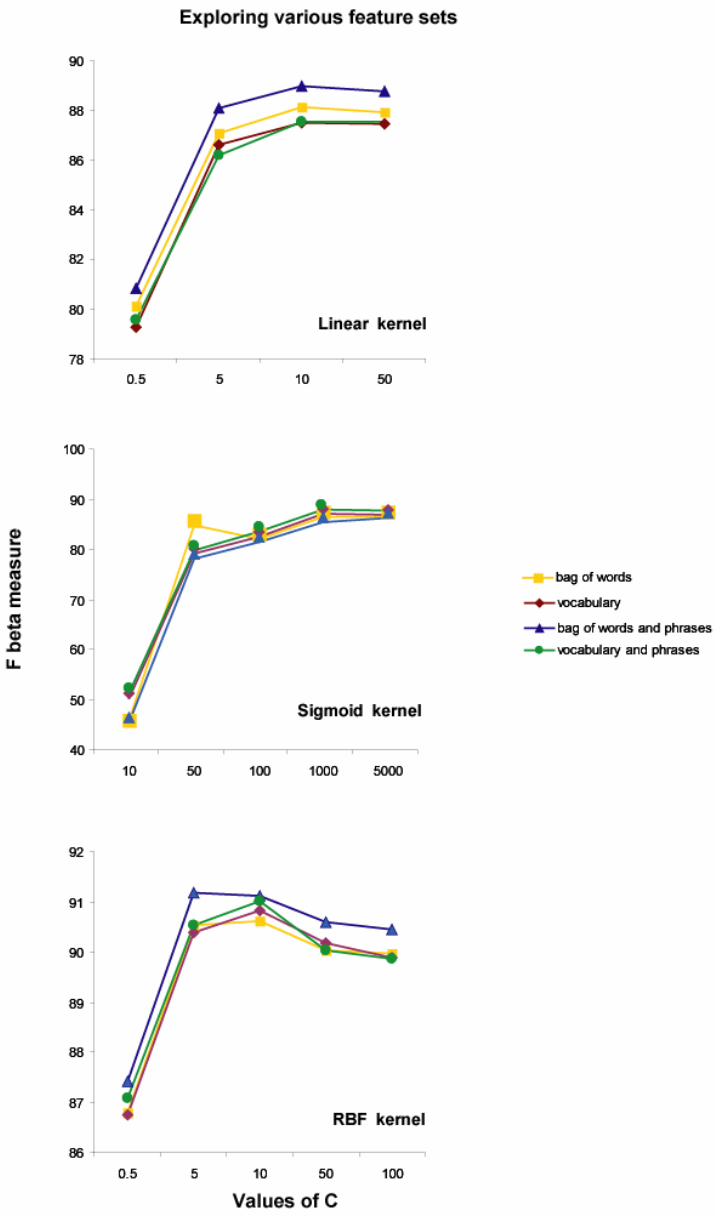

Figure 4.36 – Feature set selection for SVM learning: Performance of SVM with three different kernels at different values of C was compared using four different feature sets. Each data point is an average from four cross-validations.

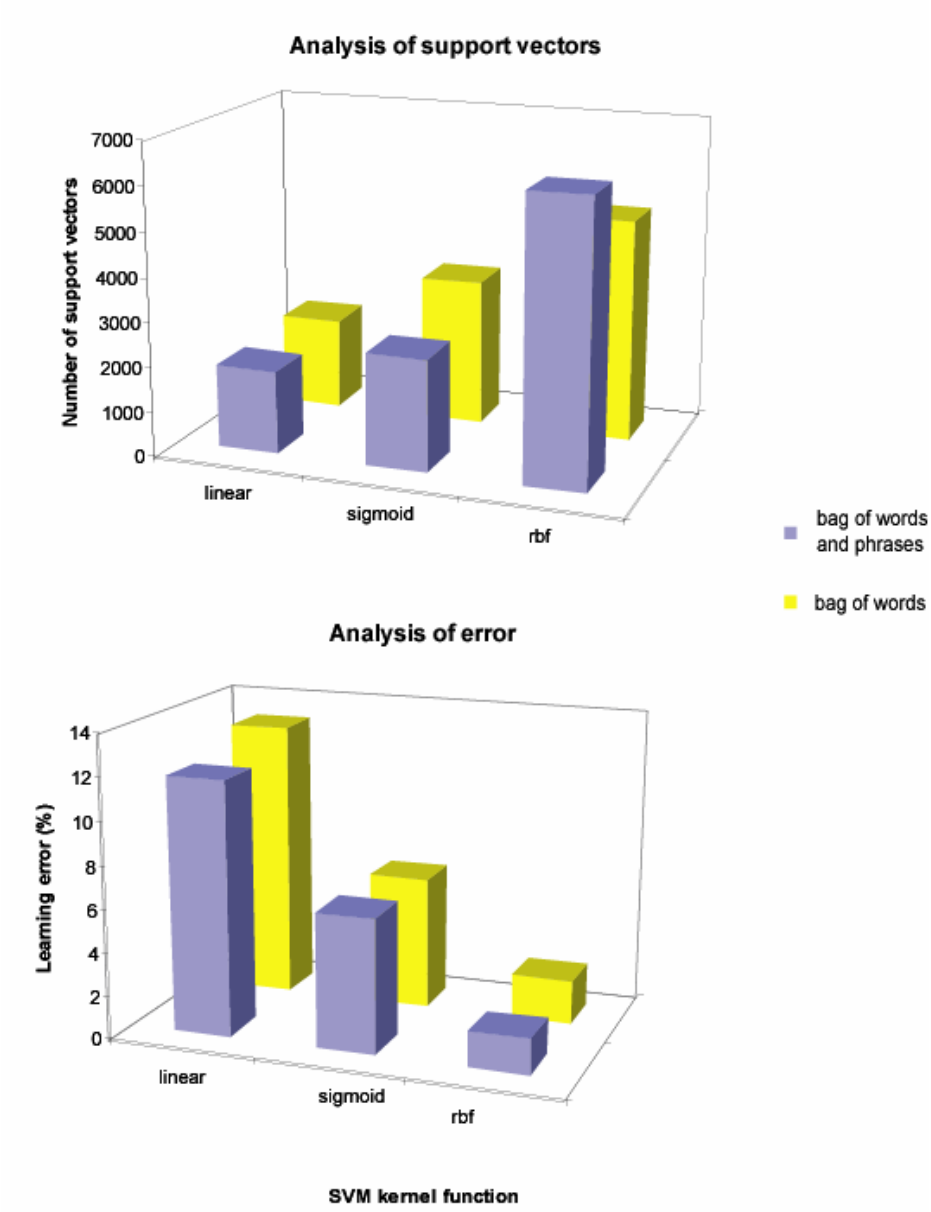

Figure 4.37 – Evaluation of SVM learning performance: Number of support vectors (top panel) and associated learning errors (bottom panel) brought about by the SVM with best learning parameters and three different kernels. Their learning performances with 'bag of words' and 'bag of words and phrases' as feature sets are shown in the figure. The C values were 10, 100 and 1000 for linear, RBF and sigmoid kernels, respectively. The gamma value of 1.5 for the RBF kernel and r value of 0.1 for sigmoid kernel were used. Each data point is an average of four cross validations.

Learning performance of the SVM

Support vectors are the training examples closest to the hyperplane and the number of support vectors utilized for deciding the margin is an indication of the complexity of an SVM model. The number of support vectors used by the SVM with linear, sigmoid and RBF kernels increased in that order (Figure 4.37). Also, for SVM with linear and sigmoid kernels the required number of support vectors decreased with the richness of input feature sets, in contrast to the SVM with the RBF kernel (Figure 4.37).

The training errors were measured using the $X_{i\alpha}$ estimators supplied with the SVM^{light} software. It gave very reliable estimates of the classifier performance on the test set. The training error is lowest when *bag of words and phrases* was used as a feature set in case of each kernel (Figure 4.37). The SVM with RBF kernel function brought the lowest training error (Figure 4.37). The polynomial functions of order more than one performed equivalently or poorly than the linear function in the experiments described above (data not shown). These results suggested that SVM with the RBF kernel was the best performing classifier (Figures 4.34, 4.36, and 4.37). Hence, the SVM with the RBF kernel with the gamma value of 1.5, C value of 10 and *bag of words and phrases* as the feature set, was used for the classification of the entire MEDLINE.

Benchmarking the classifier performance

The trained classifier identified 31,123 sentences from more than 12 million MEDLINE abstracts. A manual check for false positives resulted in retaining 20,549 sentences describing TD. This gives 66.02% accuracy to the classifier while classifying all sentences in MEDLINE. Details on the training set and the SVM training procedure are described in the Methods section.

The recall of the classifier was assessed against manual annotations of alternative splicing provided by the MEDLINE curators. All entries (5919) with the MeSH term ‘alternative splicing’ and describing the generation of physiologically relevant alternative transcript were taken from the MEDLINE 2004 database (see Methods). The classifier detected 4400 out of 5919 abstracts, resulting in a recall of 74.33%. The recall of the classifier in identifying alternative splicing in different species (human, mouse, rat and Drosophila), was also measured using manually curated AS annotations from Swissprot and MEDLINE. For each of these four species, the classifier results were compared against MEDLINE entries with the MeSH term *alternative splicing* and Swissprot entries (Bairoch and Apweiler, 2000) with the keyword *alternative splicing*. The average recall of the classifier was 61% (Table 4.32; see methods).

The abstracts missed by the SVM classifier were checked manually. In many cases the sentences (abstracts) missed by the classifier were describing alternative splicing in normal versus diseased states and they were labelled as negatives with very low confidence. However, these abstracts didn’t explicitly mention changes in gene sequence as the basis of alternative splicing. Hence, they were counted as false negatives. The F-measure while classifying all sentences in MEDLINE is 70%.

Species	MEDLINE entries			Swissprot entries		
	Total	Detected	Percentage	Total	Detected	Percentage
Human	4378	2841	64.89	2020	1364	67.52
Mouse	1537	779	50.68	1236	542	43.85
Rat	1043	600	57.52	431	305	70.76
Drosophila	277	149	53.79	331	273	82.47

Table 4.32 - Recall of the SVM classifier.

4.3.3. - Analysis of extracted sentences

The sentences describing TD extracted by the SVM classifier were divided into three different categories. Sentences in the first category are those in which the mechanism responsible for TD is present (Figure 3.41; category 1). Some of these sentences could be extracted with keywords. However, in comparison to sentences describing alternative splicing and alternative polyadenylation, those describing the use of different promoter display more variability and would be more difficult to retrieve.

Sentences belonging to the second category may describe the observation of TD but the mechanism is presumed (Figure 3.41; category 2). On the other hand, sentences without any mechanism description were also very common (Figure 3.41; category 3). For sentences in this category, the candidate mechanism was found in many cases in the full-text article when searched manually. Sentences of categories 2 and 3 may reflect practical problems while working with biological samples, lack of space in abstract or domain specific writing styles. However, the description may prove useful for elucidating the mechanism involved. For example, description in sentence 10 suggests usage of alternative first exons that may be combined with alternative promoters and alternative splicing as a plausible mechanism for sentence 11. This type of sentences may not be easily identifiable using keywords; yet exact mechanism information could be obtained by combining information extracted from the text with that coming from high-throughput data (see below).

4.3.4. - Semantic role labeling

Semantic parsing of sentences is a difficult but important task towards natural language understanding, and has immediate applications in tasks such as IE and question answering. During the task of *semantic role labeling*, for each verb in a sentence, the goal is to identify all constituents that fill a semantic role, and to determine their roles (see Introduction, supplementary material). For example, in a sentence containing the verb *express* sentence constituents may have roles such as gene name, number of isoforms, tissue-specificity and mechanism. These roles are formalized by the PAS frame semantics.

Eight frequently present semantic categories were identified in the sentences extracted from MEDLINE and performed a limited role labeling. These categories include Gene names, tissues, species, differences in structure/function of alternative transcripts, expression-specificity, number of isoforms and mechanisms

(Table 4.31). These categories were found associated with specific verbs. For example, phrases describing *mechanisms* were frequently associated with verbs like splice, express, produce, utilize, isolate, encode, and lack etc. in the context of alternative transcript expression in tissues. Similarly, phrases describing *experimental methods* were frequently seen with verbs like detect, result, report, and observe, in the sentences describing TD. Additional verbs were contain, lack, bind, show, and identify etc.

Gene and tissue names were tagged with entity taggers and all others categories were tagged using rules based on the PAS. The values for recall and precision for tagging of semantic categories were highly satisfactory (Table 4.31). The performance at the tagging boundaries was not evaluated in this study.

4.4. – Data mining of LSAT

4.4.1. - Proposing new annotations in sequence databases

There are 8133 abstracts in MEDLINE 2003 release with the MeSH term *alternative splicing* assigned to them by the annotators at the National Library of Medicine. During the information extraction step, 1536 additional abstracts describing AS events but lacking the MeSH term annotation were identified. This corresponds to a 19% increase in annotation. Moreover, new MeSH terms *alternative promoters* and *alternative polyadenylation* were proposed, for which 874 and 219 instances were extracted.

The results from the extraction step were used to provide new annotations for alternative splicing, in terms of Ensembl genes for human, mouse and rat (see Methods). The putative annotation increase observed was 20%, 52%, and 105% for human, mouse, and rat genomes, respectively (Figure 3.43). The annotation increase for the human genes was relatively little compared to that for the rat genes. Even more annotations could be obtained by manually curating extracted events that could not be automatically mapped to a sequence database entry. For instance, an additional 190 genes exhibiting tissue-specific splicing were mapped to entries in the sequence database when searching for genes exhibiting tissue-specific splicing.

4.4.2. - Quantification of the different mechanisms that lead to transcript diversity

A vast majority of the vertebrate multi-exon genes undergo alternative splicing (Johnson et al., 2003). Moreover, different promoters may control the transcription of different mRNA isoforms, which may result in directed 5’ exon inclusion/exclusion, and alternative polyadenylation signals can control the tissue specificity of alternative 3’ exons. While examples of synergy between these mechanisms are known, the extent of it is currently being explored. Differential promoter usage was found co-mentioned with alternative splicing in 14% of abstracts. A total of 19% of the abstracts providing information about alternative first exon usage also mentioned usage of different promoters. A total 17% of abstracts describing alternative polyadenylation also mention AS.

The extent to which various mechanisms are utilized for increasing transcript diversity may vary across different anatomical systems. To study this, all vertebrate tissue information was mapped to anatomical systems using the MeSH anatomy terms and the number of non-redundant events extracted for each mechanism in each system was counted (Figure 4.41; top panel).

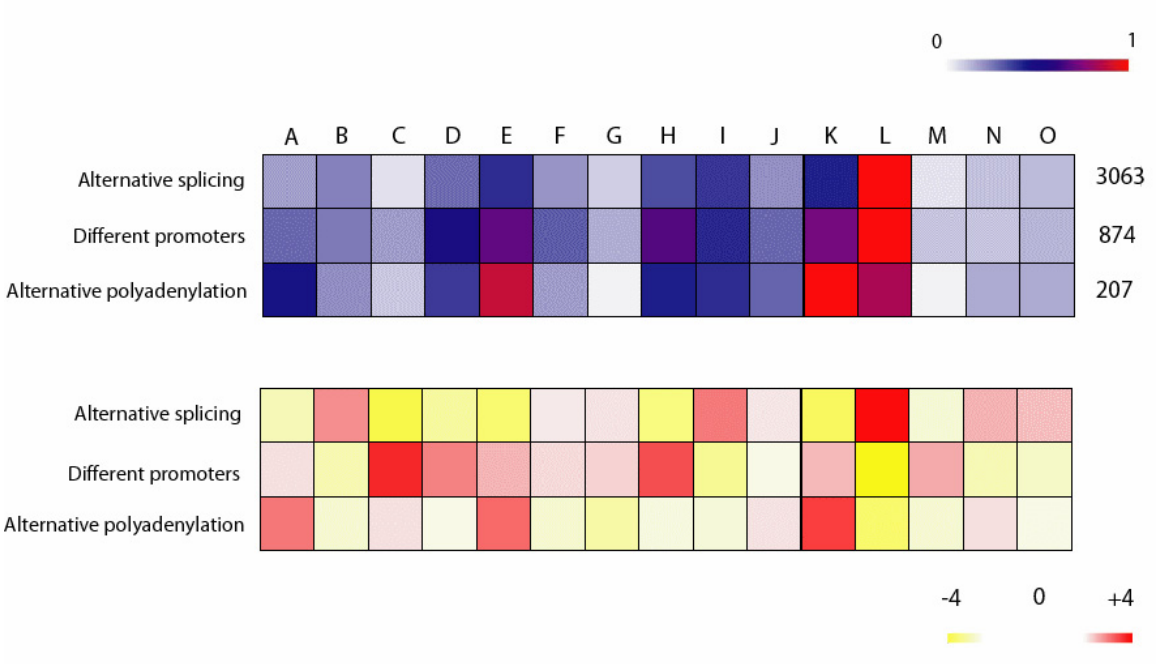

Figure 4.41 - Preference for the utilization of TD generating mechanisms across anatomical systems: Non-redundant instances of alternative splicing, differential promoter usage and alternative polyadenylation are plotted against anatomical systems in which expression was found. The color of each square in the top panel signifies the ratio of number of events detected for the system to the highest number of events within the row. Total number of non-redundant instances for each mechanism is on the left. The bottom panel shows the negative logarithm of p-values (see Methods for details). The anatomical systems are, A: cardio vascular system (sys); B: Cells; C: Connective tissues; D: Digestive sys; E: Fetal/embryonic structures; F: Endocrine sys; G: Exocrine glands; H: Genetalia; I: Immune sys; J: Integumentary sys; K: Musculoskeletal sys; L: Nervous sys; M: Respiratory sys; N: Sense regions; O: Urinal sys.

The figure shows that alternative splicing is utilized equally in most organs except in the nervous system where AS is significantly over-represented (Figure 4.41; bottom panel). Similarly, the figure shows a significant over-representation of differential promoter usage in the connective tissues and to a lesser extent in the digestive system and in the genitalia.

The knowledge about alternative promoter usage with gene names and tissues extracted in this study is the largest such collection available at present. It would provide a reliable dataset for the development of computational methods for predicting tissue-specific promoter usage.

4.4.3. - Identifying tissue specific differences in the extent of alternative splicing

With a large collection of alternative splicing events, tissue-specific differences of AS should become visible; AS has been shown to play an important role in creating functional specialization of tissues and development stages (Grabowski and Black, 2001; Yeo et al., 2004). But only a small number of instances of tissue-specific splicing are listed in the current AS databases (Thanaraj et al., 2004; Xu et al., 2002). Entries in LSAT containing the field 'specificity' were checked for information about specificity in AS and 959 such events were identified. It represented 675 AS events for pairs of tissues and 284 events where only one tissue was reported. The results contained 400 non-redundant events for 183 human genes. Moreover, a further 190 genes (not included above) from various species were mapped to Swissprot identifiers during the manual curation.

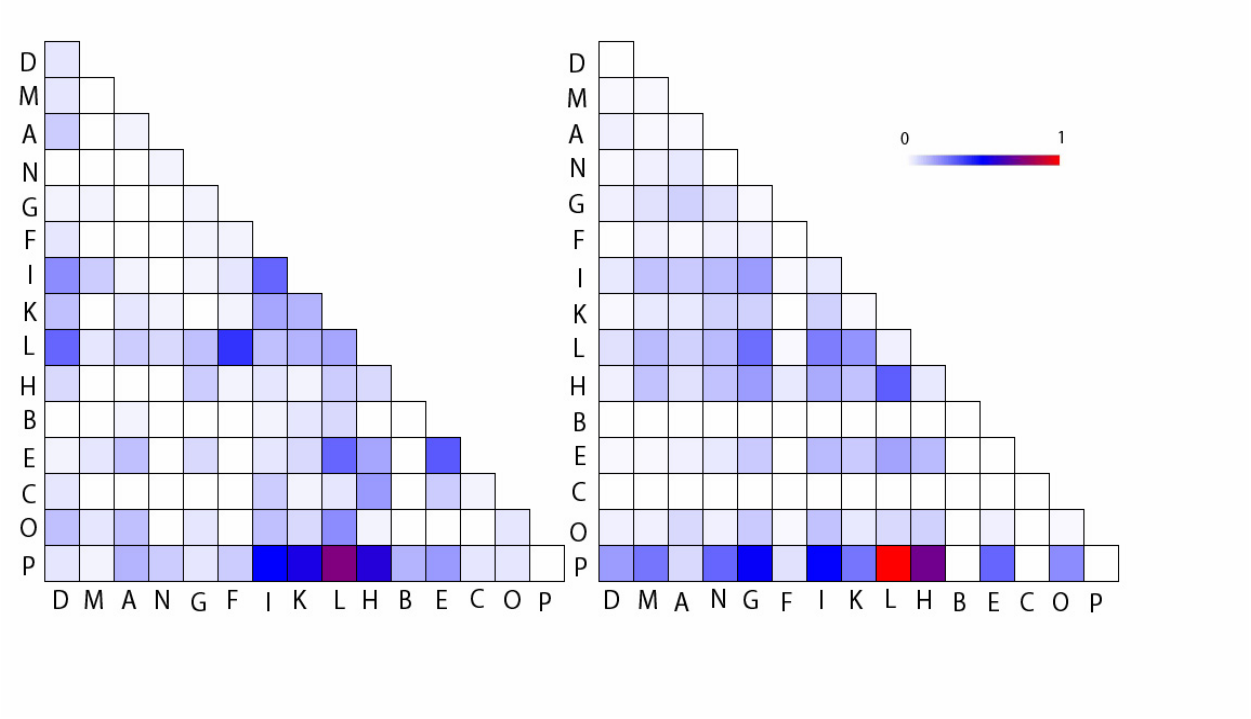

Figure 4.42 - Tissue specificity in AS: The distribution of differential/specific splicing event across different anatomical systems is shown. The instances were obtained from literature mining (left panel) and analysis of EST data ((Xu et al., 2002), right panel). Each square is colored according to the ratio between the corresponding count and the highest count within the panel. Names for systems denoted by alphabets A-O can be found in the legend of figure 4.41. P represents unique transcript.

To study the extent of tissue-specific AS, tissue/organs were mapped to respective systems as described in the previous section and plotted (Figure 4.42; left panel). The nervous system, genitalia, immune, digestive and musculo-skeletal systems showed extensive amount of tissue specificity in inter- and intra-systemic alternative splicing. These systems also showed expression of the unique AS transcripts with the nervous system showing the highest amount of unique transcripts. These tissue specific patterns of expression extracted from literature strongly overlap with the 667 tissue-specific AS events derived from the analysis of the EST data (Xu et al., 2002) for 454 human genes across 46 tissues (Figure 4.42; right panel).

4.4.4. - Assigning function to the transcripts generated by computational analysis

As mentioned before, the mechanism responsible for multiple transcripts is sometimes speculated with a limited number of experiments and the corresponding transcripts are not deposited in the Genbank. For example, work by Pisarra et al., (Pisarra et al., 2000) on human *Dopachrome tautomerase* describes two transcripts expressed in melanocytes and melanomas with a 'different carboxyl-terminus' concluding that 'dopachrome tautomerase can yield different isoforms by alternative poly(A) site usage or by alternative splicing' (Figure 4.43).

On the other hand, various methods including those based on aligning EST and other sequence data to genomic regions are currently used for detecting AS on a large scale. The function of the isoforms thus generated is largely unknown (Lee, 2003) and these transcripts are poorly annotated in sequence databases.

Using the heaviest bundling algorithm (Lee, 2003) with genomic sequence data from Ensembl (Birney et al., 2004), and transcript data from UniGene clusters (Wheeler et al., 2004) for the gene, two transcript isoforms were generated (Figure 4.43; bottom). These isoforms resembled those described in the paper and alternative splicing was detected at the 3'-region. Hence, the usage of large-scale methods may provide detailed information about the underlying event and text mining can add functional annotations to the observed transcript isoforms.

Entry: 757	Pmid 10886507
Title: Human melanocytes and melanomas express novel mRNA isoforms of the tyrosine-related protein-2 /DOPAchrome tautomerase gene: molecular and functional characterization	
Genbank:	Ensembl: ENSG00000080166.2
Refseq Report:	
Identfiers	NM_001922
species	Human
Gene Definition	Dopachrome tautomerase
Comment	PROVISIONAL
Transcript variants	
Text Extraction Data	
Gene name	tautomerase
Event	alternative poly(A) site usage or by alternative splicing
Species	human

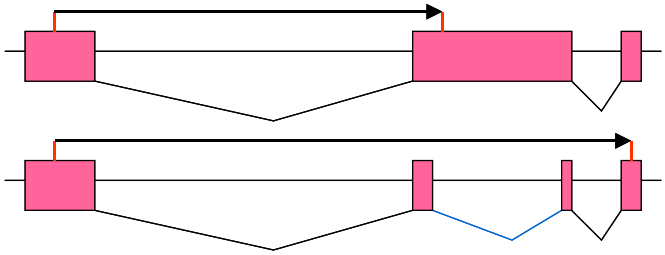

Figure 4.43 - Assignment of function using knowledge in LSAT: Figure (top) shows an LSAT entry that has very little functional annotations in sequence databases. Text extraction rules were successful in identifying gene name, tissue, and event mechanism for *Dopachrome tautomerase* gene. Multiple transcripts of the gene that resulted from utilizing alternative 3' splice site and polyadenylation signal (bottom) could be generated using SPLICE-POA (Lee, 2003). Pink squares denote the exons, black lines describe constitutive splice sites and blue lines show alternative splice sites. Black arrows show the different proteins generated due to AS.

V. - Discussion

IE and literature mining from biomedical texts was a nascent field when I started the work presented in this thesis. The number of published research articles in the field is increasing rapidly due to its potential usage in aiding knowledge discovery and assisting in analysis of data coming from high-throughput methods (Perez-Iratxeta et al., 2002; Raychaudhuri et al., 2002). The results described in this thesis flow from general analysis of the biomedical corpus to the design of a specific application and statistical data analysis. The work tackles four important tasks for applicability of IE in biomedical texts.

The first task was to check if the article full-text contains more information than abstracts of the article. It also involved the study of organization of information in the different sections such that IE attempts could be practical on the article full-text. It was found that information is heterogeneously distributed across the articles. Abstracts are the best repository of biological information followed by Introduction and Discussion sections.

The second task was to identify a suitable approach to overcome the problem of existence of multiple syntactic patterns in biomedical texts in order to build a general purpose IE method. Providing a suitable resource for general purpose IE is a non-trivial task. PAS, which provides semantic extraction frames, was found to be a naturally suitable solution for overcoming the problem of syntactic patterns in biomedical texts. Representative sentences from abstracts and full-text articles were analysed manually and with the help of linguistic parser. The concept of PAS was borrowed from the NLP in the general domain and a database of PAS (PASBio) tailored to biomedical texts was generated.

Thus, the third task was to generate a database of transcript diversity (LSAT) semi-automatically using a two-step procedure involving sentence classification and IE (semantic labeling) steps. A solution to the problem of syntactic patterns was also sought by machine learning methods. Various machine learning methods including the SVM were used for the sentence classification task for identifying sentences that describe the generation of alternative transcript isoforms in different tissues across species. A limited amount of semantic role labelling was carried out to provide sentences constituents with appropriate tags and prepare a database of experimentally verified alternative transcripts.

The last task was to apply the knowledge stored in LSAT for providing automatic annotations to abstracts in MEDLINE and sequence entries in various databases. Statistical data analysis and comparison of the knowledge in LSAT with data from high-throughput methods provided novel insights on synergy and preference of various mechanisms that generate transcript diversity.

5.1. - Analysis of full-text articles for IE

5.1.1. - Choice of the data-set

There is a clear need for utilizing the full text scientific articles for IE in biology and the primary requirements for doing it are already present. Modern computers are better suited for faster computation and storage. Regarding the source of data, electronic versions of the full-text articles are now more a rule than the exception, with initiatives towards the construction of large public repositories of such information like PubMed Central (Roberts, 2001).

Information carried by the different sections of a paper, especially between the Abstract and the rest was compared to find differences in section contents. For that, a set of full text articles with a regular section structure, namely having a defined A, I, M, R, and D sections were used as the source for the analysis. Another requirement was that of a certain homogeneity of style across the articles (for example, a similar length of the Methods section) and, since there was a great interest in the field of data mining on the detection of gene names at the time of the work, the article contents should be related to Genetics. Thus, 104 articles published in *Nature Genetics*, that comply with the AIMRD structure were chosen.

5.1.2. - The distribution of information is heterogeneous

The results showed that the distribution of information in full text articles is heterogeneous and that there is a certain correspondence of article sections with different kinds and densities of relevant data. The Abstracts were the best repository from the point of view of having many keywords in a short space, justifying previous information extraction approaches. The lack of large repositories of full text articles in contrast to the current 12 million of references in the MEDLINE database is another advantage of the Abstract approach.

However, there is much more relevant information (at least in a ratio of 1:4 regarding gene names, anatomical terms, organism names, etc.) in the rest of the article. Moreover, the information is structured enough to get important numbers of relevant keywords, but that for certain words (such as gene names), caution has to be taken regarding the context of the word.

Hence, mining of full text articles should be approached with different strategies for different sections. Beyond the Abstract, the Introduction is the best place to look for protein and gene names (and interactions) since it would likely describe the state of the art of the subject under discussion. The Discussion section, that interprets the results and puts them in context with the current knowledge, looks like the third best place for mining such information, with Methods probably as the worst place. The Results section can be problematic given its mixed nature between Methods and the rest.

5.1.3. - Introduction and Discussion are also information rich

Regarding other subjects, such as keywords about biological phenomena, and biological objects (species, tissues, diseases, etc.), again the Abstract and then the Introduction section look like the best sections to mine regarding the frequency of such keywords, but Results and especially Discussion seem better from a quantitative point of view. The Methods section is clearly appropriate for looking for technical data, measurements, and chemicals. In respect to chemicals, again, their context can be completely different in this section compared to the rest.

5.1.4. - Context matters

In brief, extraction of biological information from full text looks promising, but context must be considered. Part of this context is given by the situation of the text under analysis within the article. Therefore, tuning the extraction of information to the section is probably a good strategy, and for particular tasks some sections should be avoided.

This work also suggests a simplistic annotation that constitutes tagging a fragment of an article as belonging to a characteristic section. But further tagging using markup codes in XML style (St. Laurent, 2000) identifying biological objects and concepts (under development; see for example (Ettinger, 2002) or the GENIA project (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/)) could ultimately make text mining relatively simple. One suggestion is to develop advanced interfaces for writers of molecular biology articles that would perform XML based tagging of important concepts which could be used to link a gene name with a unique and stable link to any of the existing gene sequence databases. For this to happen, the collaboration between both scientists and publishers will be very important.

5.1.5. - Related work on analysis of full-text articles

This work was first to analyze the suitability and information contents of various sections in the full-text articles for IE. A year after this work was published; there was another report on the analysis of full-text articles using keywords (Schuemie et al., 2004). The work utilizing a corpus of ~4000 full-text articles and an expanded keyword system obtained very similar results. Schuemie and coworkers also concluded that information density is highest in the abstracts, but that the information coverage in full texts is much greater than in abstracts. Their analysis of five standard sections (AIMRD) of articles showed that the highest information coverage is located in the results section. Still, 30-40% of the information mentioned in each section is unique to that section. Only 30% of the gene symbols in the abstract were found to be accompanied by their corresponding names, and a further 8% of the gene names were found in the full text. In the full text, only 18% of the gene symbols are accompanied by their gene names.

5.2. - Exploitation of sentence semantics for accurate event extraction

The keyword based analysis described above and other similar analysis reported elsewhere provide the idea of the contents of the text. However, they don't provide the exact relationships between entities as described in the literature for populating databases with structured fields. Descriptions of relationships and entire events must be extracted from the relevant sentences by appropriate patterns (templates) on the surface text around event-depicting verbs using systems such as part of speech taggers, regular expression matches and shallow parsers (Saric et al., 2004; Wattarujeekrit et al., 2004). As shown in the Introduction section, IE methods in biomedical texts also face the problems posed by multiple syntactic patterns like all other domains of written language. Moreover, information extraction from complex sentence in biomedical texts requires deeper knowledge of sentence semantics.

By mapping the surface texts to semantic frames around a predicate, one may get a PAS frame for that predicate. PAS is a knowledge rich and useful intermediate structure for extraction of constituents of an event or relationships. Examples of use of full parsers and corpus based machine learning of predicate argument structures are now commonplace in the IE literature (Pradhan et al., 2004; Surdeanu et al., 2003; Tateisi et al., 2004). These approaches are helpful in overcoming problems posed by multiple syntactic patterns.

5.2.1. - Specialization of domains affects various text processing tools

The sentences found in biomedical texts are complex and technical in nature. The word usage patterns are different in biomedical texts compared to general English. For example, there are gene names called *Not* or *That*. Gene names like *A6* or *suppressor of Hairless* common in biomedical texts. Hence, various tools including the part of speech taggers, full parsers and PAS frames which are trained using a corpus of general English, have problems processing the biomedical text correctly. For example, Saric and co-workers improved the tagging performance of the TreeTagger (which is used in this thesis work) by more than 4% by re-training it on GENIA corpus (Saric et al., 2004).

Domain specific differences in verb semantics and argument usage was also noticed while defining PAS for interesting verbs from biomedical texts. Thus, PAS of many verbs in biomedical domain contain more (e.g., mutate, initiate) or less (e.g., block) arguments compared to that required for their PAS in general domain. Moreover, verbs like express occur with different semantics in biomedical texts compared to general English and others like transform require more than one PAS frames depending upon the context in which they occur. Thus, analysis and mining of biomedical texts would require tools which are trained using corpus of biomedical texts.

5.2.2. – PASBio: A database of predicate argument structures for molecular biology

As mentioned above differences in domain affect the verb semantics and argument structures. Hence, event extraction from biomedical texts would require a specialized database of PAS. PASBio was developed to fulfill this requirement. PASBio contain PAS structures for 30 verbs at this stage. In generating PASBio, verbs were chosen based on their frequency in the articles and also on their importance in a number of major event types such as gene expression, molecular interactions and signal transduction. At least one PAS frame per verb was defined, guided by WordNet senses (Miller, 1990).

In order to generate a PAS frame, diverse sentences containing a common predicate were analyzed with both the biological and the linguistic perspectives. The arguments constantly accompanying the predicate and important for completing the meaning of the sentence were proposed as core arguments. The core arguments were given numbers from 0 to n. Arg0 was reserved for the agent of the event and ArgR for the result of the event denoted by the predicate. As a rule ArgR denotes the argument describing the result of the event and multiple *ArgX* play distinct roles during the event. This condition is depicted by a formula like "verb event = (*Arg X* + verb + *Arg X*) + *ArgR*". Empirically, *ArgR* is used with transitive verbs like *skip*, *delete, mutate* related to abnormal biological phenomenon. Apart from the role of agent which was reserved for gene/gene products, other roles in common with the general domain such as instrument, source, and location were also frequent in biomedical texts. However, predicate usage and their argument sets differ considerably between biomedical and general English.

Construction of PAS frames by expert analysis is a time-consuming process. It is possible to approach the task of PAS definition from a machine learning approach utilizing information from the parser, and also to follow a path of hand-built mapping rules for assisting in semantic role assignments. However, a full parser was used only to assist grammatical and linguistic aspects in this work.

5.2.3. - Utilization of PASBio

Each PAS frame in PASBio provides a set of semantic relationships between arguments participating in an event and a verb conveying the event. Domain-specific PAS frame definitions have valuable uses in several applications. Although, the main focus while developing PASBio was reliable event extraction in the molecular biology domain, any information processing application that requires semantic understanding of a sentence will be able to take advantage of this knowledge. For example, machine translation (MT) that requires encoding a surface sentence of a source language into a language independent logical form of clause meaning, and then generating from this logical representation a surface sentence in a target language. PAS would provide such a logical representation in MT (Hajic et al., 2004; Han et al., 2000). In the case of a text summarization application, PAS frames could be employed as the basic unit of a discourse representation, before being summarized (Marcu, 2000). PASBio is available

online for the wider research community in the molecular biology domain for exploitation in such applications.

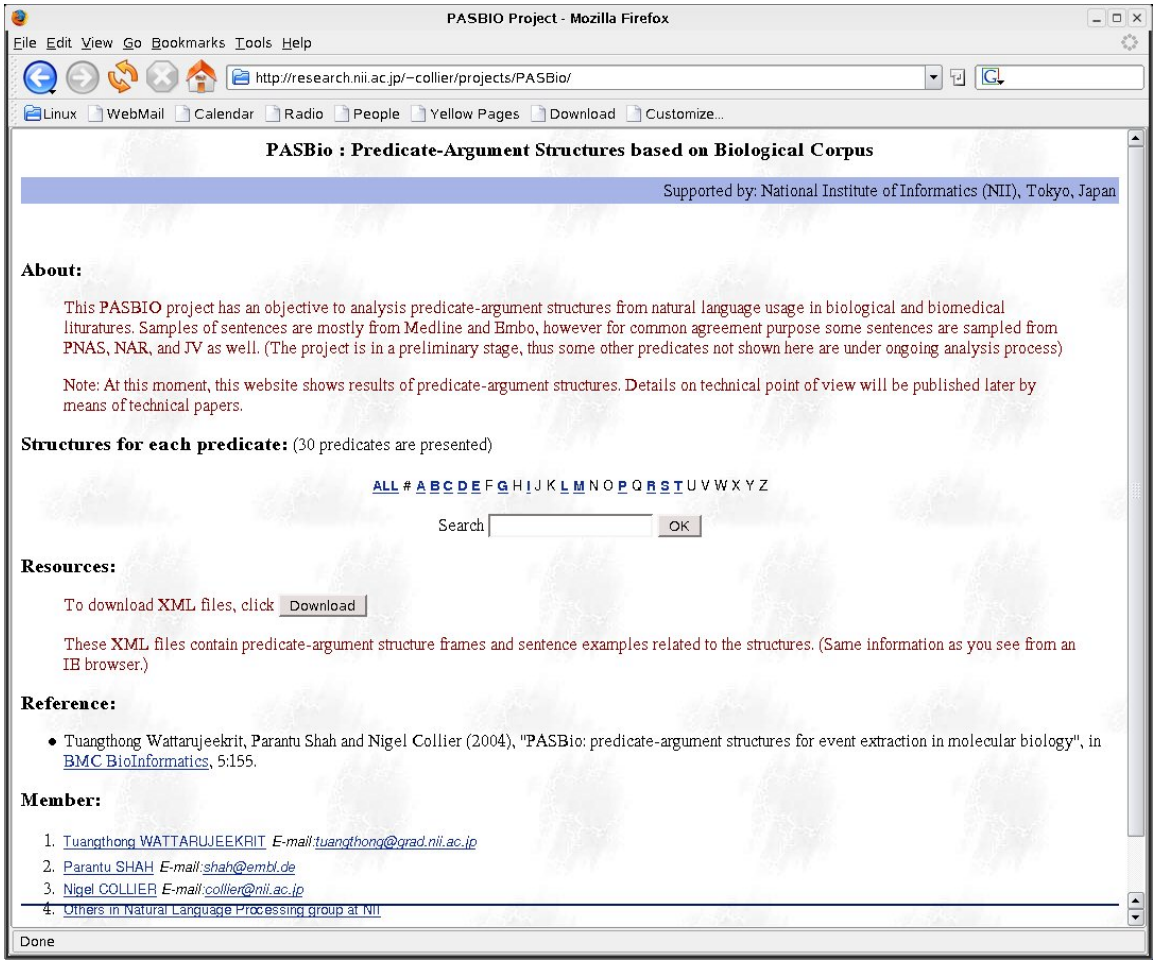

Figure 5.21 - PASBio – a database of predicate argument structures. The database was constructed for event extraction in biomedical texts and is available at http://research.nii.ac.jp/~collier/projects/PASBio

Role of PAS in a complete extraction system

With respect to a complete event extraction system in molecular biology, PASBio takes on the role of a reference source providing annotated training examples (corpus) for machine learning. The utilization of a PAS frame involves four stages: (1) construction of a semantic lexicon; (2) annotation of frame elements in sentences using the knowledge in PASBio; (3) automatically transforming sentences from surface forms to logical forms; (4) extracting the semantic relationships from the logical form and integration of the resultant interpretation within the event extraction framework. The work of Surdeanu et al. that utilized PAS defined for the newswire domain to extract market change events provide an excellent description of an IE system that makes use of a corpus annotated with PAS elements (Surdeanu et al., 2003).

5.2.4. - Related work on Information Extraction from biomedical texts

There are many initiatives for event extraction from the biomedical literature. Most of them utilize MEDLINE abstracts. Most of these approaches can be summarized into two sets. The first set of methods use regular expressions and rely on syntactic patterns. These methods may use statistical models of the surface words (Donaldson et al., 2003; Marcotte et al., 2001), rules of the sentence elements' precedence order (Blaschke et al., 1999), shallow knowledge like part of speech tags, syntactic roles of constituents (Ono et al., 2001; Pustejovsky et al., 2002), gene/protein name dictionaries and domain knowledge (e.g. a template slots for the particular event) about the events they intend to extract (Rindflesch et al., 2000; Sekimizu et al., 1998). A template used in this research group consists of only a simple set of slots for a simple predicate (i.e. the predicate relating only two arguments: subject and object) and only a shallow notion of the predicate-argument structure has been utilized (i.e. considering one argument as subject and another as object, but not considering semantic roles).

Methods in the second set, take into account a large number of linguistic and deeper semantic aspects. For example, the MedScan system (Novichkova et al., 2003) is composed of two components: an NLP engine deducing the semantic structure of a sentence, and a configurable IE component to validate and interpret results produced by the NLP engine, in order to achieve a flexible and efficient IE system. However, like many other proposed systems the semantic interpretation module of MedScan is still under development and not precisely specified. Recently, another research group (Tateisi et al., 2004) reported the aim of annotating a biological corpus with semantic knowledge in the form of PAS. This work is also at an early stage. However, these examples show the importance of predicate-argument frames and the semantics lying therein as a key knowledge for IE in the molecular biology domain. Thus, the NLP approaches, whether a deep notion of predicate-argument relations is taken (Novichkova et al., 2003) or a shallow notion (Rindflesch et al., 2000; Sekimizu et al., 1998), do require a reference resource of PAS frame for each predicate. In this respect, PASBio's description of PAS frame for each predicate will be a useful complement to other approaches.

5.3. – Generating a database semi-automatically with a two-step procedure

5.3.1. – Description of LSAT

LSAT (Literature Support for Alternative Transcripts) was generated semi-automatically using a two-step procedure (Figure 5.31). The first step was retrieving all the sentences describing TD from MEDLINE (sentence classification) and the second step was to extract the information (role labeling) from the sentences provided by the classifier.

Since LSAT was generated from MEDLINE, information associated with LSAT entries is experimentally verified. Entries in LSAT are divided in three parts (Figure 4.32). The first part contains literature data including title and abstract text of identified MEDLINE entry. The second part contains links

to various sequence databases like Swiss-Prot, Refseq, GenBank and Ensembl. The third part contains information including gene names, species, experimental methods, and mechanisms generating alternative transcripts, extracted from the sentences identified by the SVM classifier.

In total LSAT contain 9,503 instances of event mechanisms from as many abstracts and 5,028 instances of tissues with associated gene names. There are 3,063, 874, and 207 non-redundant instances of AS, differential promoter usage (DP), and AP associated with genes and tissues extracted by entity taggers. The information about alternative promoter usage linked with specific gene names and tissues extracted in this study is the largest such collection available at present. It would provide a reliable dataset for the development of computational methods to predict tissue-specific promoter usage. Moreover, LSAT can be searched using identifiers from Swissprot, Refseq, Genbank, and Ensembl, apart from gene names, species and event mechanisms. LSAT data are also available free for download.

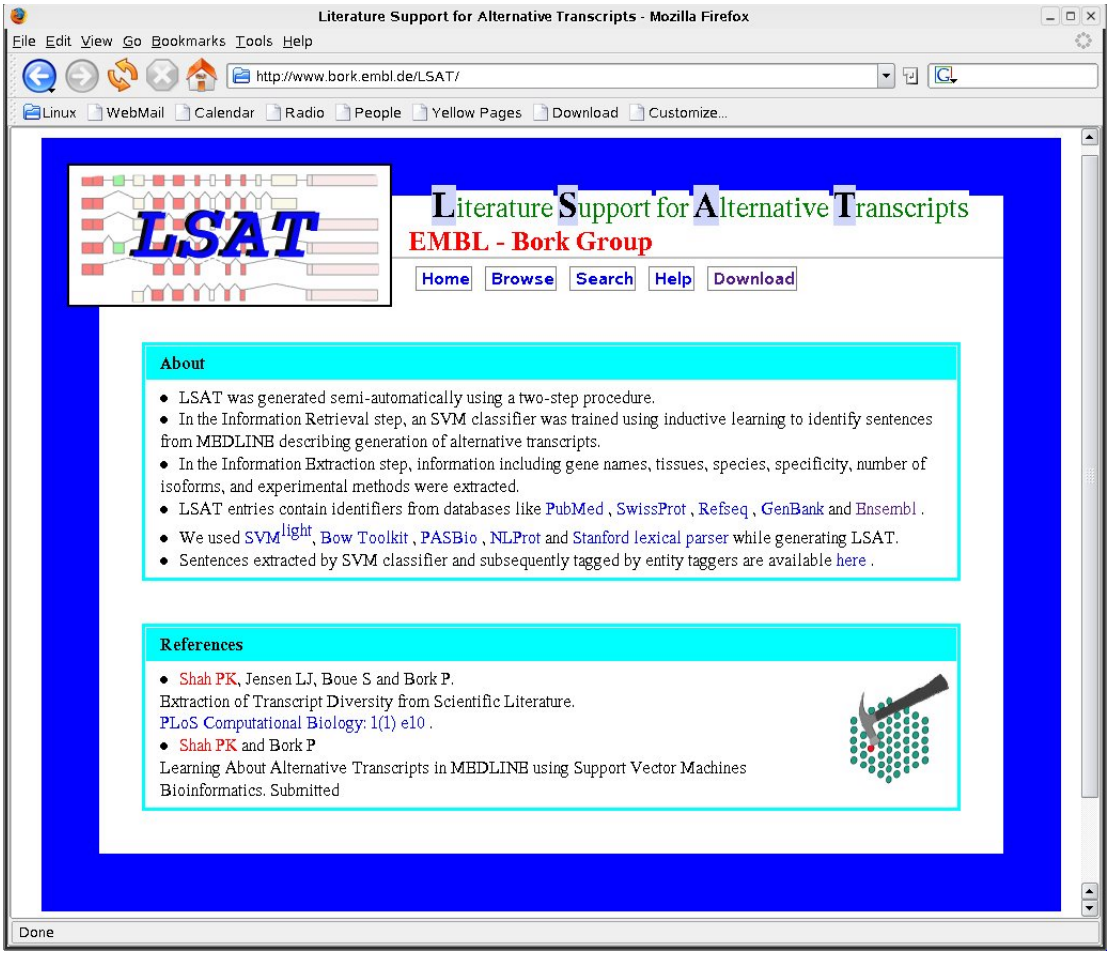

Figure 5.31 – A database of transcript diversity: Snapshot of LSAT homepage at http://www.bork.embl.de/LSAT/.

5.3.2. - Retrieving event describing sentences using text categorization methods

Why sentence classification should be attempted

As discussed above, biomedical texts utilize domain specific predicates and argument sets of other commonly used predicates are often changed. Thus, attempts for the assignment of the predicate argument relationships in biomedical tools using standard NLP tools suffer from low values of precision and recall (Wattarujeekrit et al., 2004). Efficient and accurate parsing of biomedical texts is not within the reach of current parsers (Shatkay and Feldman, 2003). Standard methods are computationally expensive to use and are trained on English texts from the newswire domain. Thus, full parsing attempts could be impractical when applied to a large database like MEDLINE. Hence, any practical event extraction task should be preceded by the identification of the event-containing sentences. This binary classification step would constrain the number of predicates, giving a better idea of the semantic roles of their arguments and reduce the computational demands. The classification would also help prioritising the predicates for the predicate argument analysis in these early days of event extraction for generating event-specific databases. Moreover, it would also help the consequent IE step to achieve higher precision.

Why SVM are superior for the sentence classification task

SVM followed by the maximum entropy classifier proved to be better than other sentence classification methods using bag of words as the input feature set. This feature set was the simplest and didn't contain any derived (e.g. synonyms) features. Hence, the method inherently good at feature selection would outperform the others in this set up. SVM benefit from this scenario as they learn classification using boundary examples (support vectors) and perform classification irrespective of the total number of input features. The use of a large training set constructed by hand was also beneficial for statistical pattern learning by the SVM. The SVM with a RBF kernel outperformed SVM with linear, sigmoid and polynomial kernel functions. The SVM with RBF kernel maps input features to infinite dimensional hyperspace thereby allowing the separation of positive examples from the negatives with maximum margin and results in better classification. Similarly, maximum entropy classifier can utilize the presence of frequently occurring phrases like 'alternative splicing' or 'multiple mRNA transcripts' for learning classification and therefore give better performance than naïve Bayes classifier (and other methods) where the underlying assumption about the word independence is disadvantageous (see Appendix). Such a trend in classification performance has been reported for the text categorization task (Dumais et al., 1998; Yiming Yang, 1999).

SVM classifiers should be utilized for extracting biological information

For the SVM with the RBF kernel, values of gamma=1.5 and C=10 led to the optimal classification performance. When the trained classifier was applied to all sentences in MEDLINE, it achieved a recall of 74.33% and a precision of 66.02% (F-measure of 70%). The average value of recall decreased to 61%, when the benchmark was carried out using Swiss-Prot annotations for the genomes of human, mouse, rat and Drosophila.

Thus, the SVM classifier was able to learn multiple patterns present in the training set while handling a relatively large amount of features and provided good values for precision and recall. This generalization performance makes SVM an attractive choice for extracting biological events from text data. Moreover, this performance was obtained over a very large repository of biomedical texts and there was no need to define any rules or pre-selecting a subset of MEDLINE. Thus, this classifier could also be utilized for mining of mRNA TD from the full text of articles

5.3.3. - Rule-based tagging for IE would help database curation

As discussed before, multiple syntactic patterns in the extracted sentences can be summarized into semantic patterns using PAS analysis (Wattarujeekrit et al., 2004). Hence, extracted sentences were analyzed and eight frequently present categories were identified with accompanying verbs. Frequent presence of the categories identified from the sentences is indicative of their biological importance. Indeed, manual annotation of information like gene name, species, tissue, expression-specificity, alternative exon etc. could be found in the Alternative exon database at the European Bioinformatics Institute (Thanaraj et al., 2004). The results of semantic role labelling step are deposited in the Alternative exon database.

Genes and tissues were tagged with named entity taggers and all others categories using rules based on PAS. Some of the verbs identified from sentences were not present in PASBio and hence they were analysed for PAS. The values of precision and recall of the IE step is highly satisfactory, however, it should be noted that accuracy in finding tag boundaries were not considered. Also, the recall is good for all categories, but not all eight categories are equally represented in the sentences.

It should be noted that not all extracted sentences provide all types of information. For example, gene names are present in ~70% of extracted sentences. On the other hand information about tissue-specificity was found only in 5% of the sentences, reflecting relatively fewer known examples in the literature (Table 5). However, in many cases, more than one sentence per abstract was extracted, containing information to complete the event description. Also, we have retained the identity (PMID and sentence number in the abstract) of the extracted sentences. Thus, missing information (e.g., gene name) can be obtained from neighbouring sentences using NLP techniques like discourse context and reference resolution, using MeSH terms associated with the abstracts or by searching for literature associated with (gene based) entries in sequence databases like Swiss-Prot (Bairoch and Apweiler, 2000), RefSeq (Pruitt and Maglott, 2001) or GenBank (Benson et al., 2004).

5.3.4. - Rule based versus semantic role labeling using machine learning

Event extraction has been carried out traditionally by writing rules for filling pre-defined templates based on syntactic patterns exhibited by event-containing sentences (Hoffmann et al., 2005). However, researchers in general domain as well as in biomedical NLP are moving towards use of predicate argument frames (Daraselia et al., 2004; Gildea and Jurafsky, 2002; Novichkova et al., 2003; Surdeanu et al., 2003; Tateisi et al., 2004). Please see the discussion above on event extraction from biomedical texts and utilization of PASBio for the same (Wattarujeekrit et al., 2004). The work of Surdeanu et al. and Pradhan et al discusses use of PAS based role tagging/labelling using machine learning (Pradhan et al., 2004; Surdeanu et al., 2003).

Machine learning of semantic role labelling is gaining importance and many community wide efforts are organized for general English (e.g. CoNLL-2005 task defined at http://www.lsi.upc.edu/~srlconll). As noted before, domain-specific corpus is required for text-processing tasks in biomedical texts. Thus, the limiting step for the learning role-labelling for the biomedical NLP is the availability of a comprehensive database of predicate argument structures and an annotated corpus. A database of predicates common in biological texts was prepared in this work and new predicates are regularly being added to it (Wattarujeekrit et al., 2004). Sentences identified by the SVM classifier and tagged with the IE step could be used as a learning corpus for semantic role labelling task for biomedical texts. In other words, the task of extracting information including gene names, species, experimental methods, mechanism could be seen as a machine learning task of mapping surface text of the sentence to its logical form using predicate frames for accurate tagging of semantic roles (Pradhan et al., 2004; Surdeanu et al., 2003).

5.3.5. - Related work on relationship/event extraction

Craven et al developed systems to distinguish fact-bearing sentences from “uninteresting” sentences for identifying protein subcellular localization and gene-disorder association. Their naïve Bayes classifier that doesn’t use grammatical rules achieved a precision of 77% and a recall of 30%. The classifier that used grammatical rules and parsing of sentences achieved a higher precision (92%) but a lower recall (21%). An important result of these experiments is the actual comparison of classifiers to a baseline method, which uses co-occurrence alone. The latter method decides that a sentence reports a “subcellular localization” fact if both a protein name and a localization word occur in it. This simple method, which is currently most popular in the context of literature mining in Bioinformatics, reaches a much lower precision than the classifiers (about 35% precision at recall 30% and 45% precision at recall 21%). The co-occurrence based method can reach a higher level of recall (~70%) without losing much in precision (~40%). However, at this recall level, a naïve Bayes classifier with a noisy ‘OR’ combination still reaches a somewhat higher level of precision (~45-50%). The study suggests that classifiers at the sentence level have the potential to improve the precision of IE, in the biomedical context, over co-occurrence-based methods. An SVM classifier with the RBF kernel was also used by curators of BIND for their Pre-BIND

and Textomy system (Donaldson et al., 2003). They combined information retrieval with information extraction to assist in recovering protein-protein interactions and found interaction information in 60% of the extracted abstracts. Saric and co-workers extracted gene regulatory network from abstracts related to baker's yeast with an accuracy of 83-90% without providing any estimated for recall (Saric et al., 2004).

5.4. - Analysis and integration of text-mining data to present knowledge

5.4.1. – Automated MeSH term assignments to Abstracts

More than 13,000 instances of event mechanisms and ~16,000 instances of gene names were present in the sentences extracted using the SVM classifier and these were stored in LSAT. Utilizing the knowledge in LSAT resulted in 19% increase in MeSH term annotation seen while comparing the tagged events to annotations provided by the annotators at NLM. This increase in MeSH term annotations it self justifies the need of IE approaches. Abstracts were also assigned new keywords (MeSH terms) of alternative promoters and alternative polyadenylation.

5.4.2 - Function annotation using text-mining

The putative increase in gene annotation was 20%, 52%, and 105% for genomes of human, mouse, and rat, respectively. These results perhaps reflect the extent of manual curation efforts that are underway for the curation of different genomes. The annotation increase for human genes was relatively little compared to that for the rat genes because a total 3438 genes are already annotated in Swissprot and RefSeq for AS in human, whereas only 342 genes are annotated for AS in rat. An additional 190 genes were mapped to swissprot. Also, functional assignments were provided for de-novo generated transcripts. Hence, the increase in the annotation reflects the usefulness of the current approach and emphasizes the need for automated methods to speed up the process of database curation.

5.4.3. - Transcript diversity generating mechanisms, synergy and preference

The synergy between different TD-generating mechanisms was explored using the knowledge stored in LSAT. The results indicate that 14% of abstracts mentioning differential promoter usage also mentioned AS and for alternative polyadenylation the co-mentioning with AS was 18%. These numbers may be the lower bound of the mechanism synergies as we are just entering the high-throughput era.

Text-mining results showed an over-representation of alternative splicing events in the nervous system. This results are inline with many EST based studies (Xu et al., 2002; Yeo et al., 2004) that report highest number of AS in the nervous system, as did earlier experimental studies (Mirnics and Pevsner, 2004). EST-based studies (Yeo et al., 2004) also suggested that genes in liver (digestive system) and testis (genitalia) show distinct pattern of splicing with alternative exons. Text-mining results indicate that these transcripts may show these different patterns of splicing in combination with different promoter regions. This conclusion seems plausible as alternative splicing of first exons is influenced by alternative promoter

regions in at least 19% of cases (results in section above; (Zavolan et al., 2003)) and should be explored further.

VI. Conclusions

1. - There is a clear need for utilizing full-text articles for IE in biology. IE from the full-text articles is required as the distribution of information in full text articles is heterogeneous and there is certain correspondence of article sections with different kind and density of relevant data.

2. - Abstracts of biomedical scientific articles are the best repository from the point of view of keyword density and availability, justifying IE approaches where only Abstracts are utilized. However, there is much more relevant information in the rest of the article, specifically in Introduction and Discussion sections. Moreover, the information is structured enough to get large numbers of relevant keywords

3. – The analysis of sentences from abstracts and full-text articles of biomedical texts demonstrates a clear need for the utilization of semantic knowledge for accurate information extraction. PAS frames provide a semantic extraction template for an event depicting predicate. Also, the semantic knowledge residing in PAS frames will help the extraction process to overcome the problem of multiple syntactic patterns.

4. - Predicate usage in biomedical texts is domain specific and hence a domain-specific PAS resource is required for accurate IE. Utilization of PAS will also allow building of a general purpose IE system for biomedical texts. The PASBio database generated as a part of this work promises to serve this and other functions (availability: http://research.nii.ac.jp/~collier/projects/PASBio/).

5. - Generation and regulation of alternative transcripts is an important event for functional diversity and evolution of eukaryotes. A database of alternative transcripts (LSAT) was generated semi-automatically using a composite procedure containing sentence identification and information extraction steps. LSAT is available at http://www.bork.embl-heidelberg.de/LSAT/.

6. - Support vector machines followed by the maximum entropy classifier outperformed other sentence classification methods. SVM with radial basis kernel function generalized well; they are the best classifiers of the text data. Machine learning of sentence classification also allowed circumventing the problem of multiple syntactic patterns. Both, the sentence classification and the information extraction steps achieved a good F-measure in the benchmarking process.

7. – LSAT is knowledge rich and knowledge residing in LSAT could be utilized for automated assignment of the MeSH terms, and function annotations to gene entries in sequence databases and de novo generated alternative transcripts.

8. – The data-mining of LSAT also allowed hypothesis testing. The results of data mining and comparison to EST data suggested that alternative splicing may be the preferred mechanism for generating alternative transcripts in the nervous system. Thus, text-mining not only assisted in analysing the data from other sources but also acted as a stand-alone data source.

VII. Supplementary Material

Appendix A

1. Glossary of Terms

Natural language processing

Natural language processing is concerned with all aspects and stages of converting spoken, handwritten or printed text from raw signal to information that can be used by either humans or automated agents. In the context of Bioinformatics, we are concerned only with the printed text that is already stored in machine accessible format and therefore concentrate on common text processing operations as used by typical text mining systems. These include the tokenization and zoning tasks, part of speech tagging and (shallow) parsing In this section I introduce general techniques from natural language processing and then proceed to more specific area of information extraction.

Tokenization

The first step in text analysis is the process of breaking the text up into its constituent units—or tokens. Tokens may vary in granularity depending on the particular application. Consequently, tokenization can occur at a number of different levels: the text could be broken up into chapters, sections, paragraph, words, syllables, or phonemes. The most common form of tokenization in mining systems is the fragmentation of text into words and sentences. The main challenge of fragmentation at the sentence boundaries is distinguishing between a period that signals an end of sentence and a period that is part of a previous token like the shorthand Mr., Dr., etc.

Part of speech tagging.

Part-of-speech tags are a set of word-categories based on the role that words may play in the sentence in which they appear. *Part of Speech (*PoS*)* tagging is the annotation of words with the appropriate PoS tags, based on their context within the sentence. PoS tags convey information about the semantic content of a word. Nouns usually denote tangible and untangible entities while propositions express relationships between entities. While sets of tags may vary most part-of-speech tag sets make use of same basic categories. The most common tags include: Article, Noun, Verb, Adjective, Preposition, Number and Proper Noun.

Several approaches exist to PoS tagging. The methodologies used in training the taggers include decision trees; hidden markov models (HMMs) or rule-based tagging. In order to estimate the model parameters the HMM tagger undergoes a training phase, using an annotated corpus such as the WSJ corpus in the Penn Tree Bank. Using tri-gram model, HMM-based taggers have achieved 94-96% accuracy in held out tests.

Other taggers may use decision trees, a searching based approach for learning and performing POS tagging. Rule based approaches rely on rules that use contextual information to assign tags to unknown or ambiguous words. These rules are also known as context frame rules. In addition to contextual information, many rule based taggers use morphological information to aid the disambiguation process. For example, if an ambiguous/unknown word ends with an *ing* suffix and it is preceded by a verb, it may be tagged as a verb. Another correct tagging of words can be obtained from orthography such as capitalization or punctuation. All approaches that are utilized for POS tagging may also be utilized for named entity tagging in biomedical texts.

Parsing and shallow parsing.

Parsing is the process of determining the complete syntactic structure of a sentence or a string of symbols in a language. A parser usually takes as its input a sequence of tokens that were extracted from the original text by a lexical analyser. The output from the parser is typically an abstract syntax tree, whose leafs correspond to the individual words (lexemes) in the text, and whose internal node represent syntactic structures identified by grammatical tags, such as Noun, Verb, Noun phrase, Verb phrase, etc. Efficient and accurate parsing of unrestricted text is not within reach of current techniques. Standard algorithms are too expensive to use on very large corpora and are not robust enough.

A practical alternative is shallow parsing. This is a coarser process of breaking documents into non-overlapping word sequences or phrases, such as syntactically related words are grouped together. Each phrase is then tagged by one of a set of predefined grammatical tags such as Noun Phrase, Verb Phrase, Prepositional phrase, Adverb Phrase, Conjugative Phrase, and List Marker. Shallow parsing has the benefit of both speed and robustness of processing, which comes at the cost of compromising the depth and fine-granularity of the analysis. Shallow parsing is generally useful as pre-processing step, either for bootstrapping –extracting information from corpora for use by more sophisticated parsers—or for end-user application such as information extraction. Shallow parsing allows the identification of relationships between the object, the subject and any other spatial or temporal phrases within a sentence.

Each of the NLP task described above and others including text categorization, named entity extraction could be learned from a suitable corpus labelled with appropriate labels. Support vector machines were compared to other text categorization methods and they are described below.

Appendix B

2. Machine learning

Machine learning draws on concepts and results from many fields, including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory. Hence, there are different learning concepts and a variety of learning methods depending upon the exact learning task. Learning can be defined in many ways. Some of the common definitions are:

- Learning denotes changes in a system that enables it to do the same task more efficiently the next time. – Herbert Simon
- Learning is constructing or modifying representations of what is being experienced. – Ryszard Michalski
- Learning is making useful changes in our minds. – Marvin Minsky.

Machine learning of Well-posed problems

A computer program is said to learn from experience *E* with respect to some class of Tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

Unsupervised learning

In unsupervised or self-organized learning there is no teacher to oversee the learning process. In other words, there is no specific example of the function to be learned by the method. **Clustering** is an unsupervised task. Clustering algorithms divide data into natural groups (clusters). Instances in the same cluster are similar to each other, they share certain properties.

Supervised learning

An essential ingredient of supervised learning is the availability of an external teacher. One may think of teacher as having knowledge of the environment that is represented by a set of input examples and their class labels. During the training session, when the learning method is presented an input vector, by the virtue of the inbuilt knowledge the teacher is able to provide the learning method with a desired or target response (usually a function) for that training vector. The learning parameters are adjusted under combine influence of the training vector and the training error. The training error can be defined as the difference between the actual response of the learning method and the desired response. **Classification** requires supervised learning i.e. training data has to specify what we are trying to learn (the classes).

Support Vector Machines:

As mentioned before SVM is a learning algorithm for a linear classifier which tries to maximize the margin of confidence of the classification on the training data set (Joachims, 2001; Nello Cristiani, 2000; Vapnik, 1999). SVM were developed by Vapnik et al. based on Structural Risk Minimization principle from statistical learning theory. The idea of structural risk minimization is to find a hypothesis *h* from a hypothesis space H for which one can guarantee the lowest of error *Err(h)* for a given training sample of *n* examples. SVM learn linear threshold functions of the type

$$h(\vec{x}) = sign\{\vec{w}\cdot\vec{x}+b\} = \begin{cases} +1,.....\forall.\vec{w}\cdot\vec{x}+b>0 \\ -1,.....otherwise \end{cases}$$

Each such linear threshold function corresponds to a hyperplane in feature space. The sign function $sign\{\}$ returns a 1 for positive argument and -1 for a non-positive argument. This means that the side of the hyperplane on which an example $\vec{x}$ lies determines how it is classified by $h(\vec{x})$.

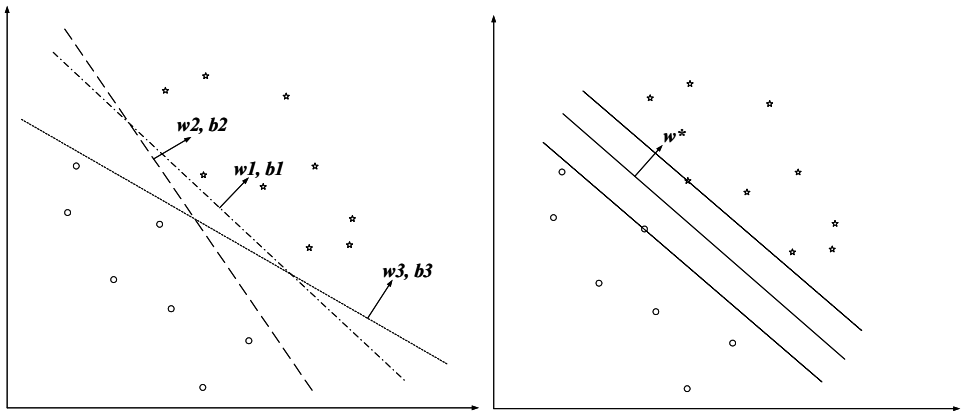

Figure 7.21 - Classification with maximum margin: The figure on the left shows how multiple hyperplanes could be defined for binary classification. SVM tries to define a hyperplane with maximum margin separating the training examples.

The way SVM function can be explained as follows. Let us assume that the training data can be separated by at least one hyperplane h'. This means that there is a weight vector $\vec{w}'$ and a threshold b', so that all positive examples are on one side of the hyperplane while all negative examples lie on other side. This is equivalent to requiring $y_i[\vec{w}'\cdot\vec{x}_i+b']>0$ for each training example $(\vec{x}_i, y_i)$. In general, there can be multiple hyperplanes that separate the training data without error (figure X). From these separating hyperplanes the support vector machine chooses one with the largest margin δ, as shown by the hyperplane $h(\vec{x}^*)$ in the figure. The margin δ is the distance from the hyperplane to closest training examples. For each separable training set, there is only one hyperplane with maximum margin. The examples closest to the hyperplane are called support vectors (marked with circles in FigX). They have a distance of exactly δ.

A remarkable property of SVMs is that they can be transformed into non-linear learners. In principle, the approach used is as follows. The attribute vector $\vec{x}_i$ are mapped into a high-dimensional feature space X' using a non-linear mapping $\Phi(\vec{x}_i)$. The SVM then learns the maximum-margin classification rule in feature space X'. Despite the fact that the classification rule is linear in X', it is non-linear when projected into the original input space. In general such a mapping $\Phi(\vec{x})$ is inefficient to compute. But in practice it is sufficient to compute dot-products in the feature space, i.e. $\Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$. For special mappings $\Phi(\vec{x})$ such dot-products can be computed very efficiently using kernel functions $\kappa(\vec{x}_1, \vec{x}_2)$. If a function $\kappa(\vec{x}_1, \vec{x}_2)$ satisfies mercer's theorem, it is guaranteed to compute the inner product of the vectors $\vec{x}_1$ and x_2 after they have been mapped into a new "feature" space by some non-linear mapping Φ:

$$\Phi(\vec{x}_1) \cdot \Phi(\vec{x}_2) = \kappa(\vec{x}_1, \vec{x}_2)$$

For example, depending on the choice of kernel function, SVMs learn polynomial classifiers, radial basis function (RBF) classifiers, or two layer sigmoid neural nets.

$$K_{poly}(\vec{x}_1, \vec{x}_2) = (\vec{x}_1 \cdot \vec{x}_2 + 1)^d$$

$$K_{rbf}(\vec{x}_1, \vec{x}_2) = \exp(-\gamma(\vec{x}_1 - \vec{x}_2)^2)$$

$$K_{sigmoid}(\vec{x}_1, \vec{x}_2) = \tanh(s(\vec{x}_1 \cdot \vec{x}_2) + p)$$

TFIDF classifiers and Rocchio algorithm:

This type of classifier is based on the relevance-feedback algorithm originally proposed by Rocchio for vector-space retrieval model. It has been extensively used for text classification. There are many different implementations of the algorithm depending on the word weighting method, the document length normalization and the similarity measure. The most popular algorithm utilizes "tf" word weights, document length normalization using Euclidian vector length and cosine similarity.

The algorithm utilize following representation of documents. Each document d is presented as a vector $\vec{d}$ = (d_1, .. , $d_{|F|}$) so that the documents with similar content have similar vectors (according to a fixed metric). Each element d_i represents a distinct word w_i. d_i for a document d is calculated as a combination of the statistics TF (w_i, d) and DF (w_i). The term frequency TF (w_i, d) is the number of times word w_i occurs in document d and the document frequency DF (w_i) is the number of documents in which word w_i occurs at least once. The inverse document frequency IDF (w_i) calculated from the document frequency.

$$\text{IDF}(w_i) = \log\left(\frac{|D|}{DF(w_i)}\right)$$

Here, |D| is the total number of documents. Intuitively, the inverse document frequency of the word is low if it occurs in many documents and is highest if the word occurs only in one. The weight d_i of the word w_i in the document is then

$d_i = TF(w_i, d) * IDF(w_i)$

This word weighing heuristic says that a word w_i is an important indexing term for document d if it occurs frequently in it (the term frequency is high). On the other hand, words which occur in many documents are rated less important indexing terms due to their low inverse document frequency.

Learning is achieved by combining document vectors into a prototype vector $\vec{c}_j$ for each class C_j. First, both the normalized document vectors of the positive and negative examples for a class are summed up. The prototype vector is then calculated as a weighted difference of each.

$$\vec{c}_j = \alpha \frac{1}{|C_j|} \sum_{\vec{d} \in C_j} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - C_j|} \sum_{\vec{d} \in D - C_j} \frac{\vec{d}}{\|\vec{d}\|}$$

Here, α and β are parameters that adjust the relative impact of positive and negative training examples. C_j is the set of training documents assigned to class j and $\|\vec{d}\|$ denotes the Euclidian length of a vector $\vec{d}$.

Naïve Bayes Classification:

Naïve Bayes is a simple text classification algorithm for learning from labelled data. The parameterisation given by naïve Bayes defines an underlying generative model assumed by the classifier. In this model, the class is first selected according to class prior probabilities. Then, the generator creates each word in a document by drawing from a multinomial distribution over words specific to the class. Thus, this model assumes each word in a document generated independently of the others given the class.

Naïve Bayes forms maximum a posteriori estimates for the class-conditional probabilities for each word in the vocabulary V, from labelled training data D. This is done by counting frequency that word w_t occurs in all word occurrences for document d_i in class c_j, supplemented with Laplace smoothing to avoid probabilities of zero.

$$P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_t, d_i) P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_s, d_i) P(c_j \mid d_i)},$$

where, $N(w_t, d_i)$ is count of number of times word w_t occurs in document d_i, and $P(c_j \mid d_i) \in \{0,1\}$ as given by the class label.

The prior probabilities of each class are calculated in a similar fashion, counting over documents instead of words.

$$P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j \mid d_i)}{|C| + |D|}$$

At the classification time we use these estimated parameters by applying Bayes's rule to calculate the probability of each class label and taking the most probable class as the prediction. This makes use of the naïve Bayes independence assumption, which states that word occurs independently of each other, given the class of the document.

$$P(c_j \mid d_i) \propto P(c_j) P(d_i \mid c_j)$$

$$= P(c_j) \prod_{k=1}^{|d_i|} P(wd_{i,k} \mid c_j)$$

The overly-string word independence assumption causes naïve Bayes to predict extreme (nearly 0 or 1) posterior class probabilities. However, while these estimates are poor, naïve Bayes classification accuracy is typically high. This can be explained in part because classification is only a function of which class has the maximum posterior, and is not concerned with its actual value.

Classification by Expectation Maximization:

If we extend the supervised learning setting to include unlabeled data, the naïve Bayes equations presented above are no longer adequate to find maximum a posteriori parameter estimates. The Expectation-Maximization (EM) technique can be used to find local maximum parameter estimates.

EM is an iterative statistical technique for maximum likelihood estimation in problems with incomplete data. Given a model of data generation, and data with some missing values, EM will locally maximize the likelihood of the parameters and give estimates for missing values. The naïve Bayes generative model allows for the application of EM for parameter estimations.

In implementation, EM is an iterative two-step process. Initial parameter estimates are set using standard naïve Bayes from just labelled documents. Then we iterate E- and M-steps. The E-step calculates probabilistically-weighted class labels, $P(c_j \mid d_i)$ using the equation above for every unlabeled document. The M-step estimates new classifier parameters using all documents, by first two equations, where $P(c_j \mid d_i)$ is continuous as given by the E-step. We iterate the E- and M-steps until the classifier is converged. Thus, this classifier can significantly increase text classification accuracy, when given limited amount of labelled data.

Maximum Entropy classifier:

The motivating idea behind maximum entropy is that one should prefer the most uniform models that also satisfy any given constraints. For document classification, the task is to learn conditional distribution of class from documents labelled with class. More specifically, we use training data to set constraints on the conditional distribution. Each constraint expresses a characteristic of the training data that should also be

present in the learned distribution. We let any real-valued function of the document and the class be a feature, $f_i(d,c)$. Maximum entropy allows us to restrict the model distribution to have the same expected value for this feature as seen in the training data, $|D|$. Thus, the learned conditional distribution P(c/d) must have the property

$$\frac{1}{|D|}\sum_{d\in D} f_i(d,c(d)) = \sum_d P(d)\sum_c P(c\mid d) f_i(d,c).$$

In practice, the document distribution $P(d)$ is unknown and we use training data without class labels, as an approximation to the document distribution and enforce the constraint

$$\frac{1}{|D|}\sum_{d\in D} f_i(d,c(d)) = \frac{1}{|D|}\sum_{d\in D}\sum_c P(c\mid d) f_i(d,c).$$

Thus, while using maximum entropy, the first step is to identify set of feature functions that will be useful for classification. Then for each feature, measure its expected value over the training data and take this to be a constraint for the model distribution.

K-Nearest Neighbour (KNN) classifier:

KNN classifier is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidian distance or Cosine. In this classification paradigm, k nearest neighbours are computed first Then similarities of one sample from testing data to the k nearest neighbours are aggregated according to class of the neighbours, and testing sample is assigned to most similar class. One of the advantages of KNN is that it is well suited for multi-modal classes as its classification decision is based on a small neighbourhood of similar objects (i.e. the major classes). So, even if the target class is multi-modal (i.e., consists of objects whose independent variables have different characteristics for different subsets), it can still give a good accuracy. A major drawback of the similarity measure used in KNN is that it uses all features equally in computing similarities. This can lead to poor similarity measures and classification errors, when only a small subset of the features is useful for classification.

Appendix C

3. Predicates present in PASBio

Group A : same sense, more arguments
alter, begin, develop, disrupt, inhibit, initiate, mutate, proliferate, skip
Group B : same sense, less arguments
Generate, block, decrease, lose, modify
Group C : same sense, same structure
abolish, confer, eliminate, lead to, result, delete
Group D : different sense or not occur
Splice, express, truncate, translate, encode, transform, catalyse, transcribe, recognize

Table 7.31: predicates present in PASBio

VIII. References

Alphonse, E., Aubin, S., Bessieres, P., Bisson, G., Hamon, T., Lagarrigue, S., Nazarenko, A., Manine, A., Nedellec, C., Vetah, M.*, et al.* (2004). Event-based Information Extraction for the biomedical domain: the Caderige project. Paper presented at: Joint Workshop on Natural Language Processing in Biomedicine and its Application (Geneva, Switzerland).

Andrade, M. A., and Bork, P. (2000). Automated extraction of information in molecular biology. FEBS Lett *476*, 12-17.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research *28*, 302-303.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res *28*, 45-48.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. Paper presented at: 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL 1998) (Montreal).

Becker, K. G., Hosack, D. A., Dennis, G., Jr., Lempicki, R. A., Bright, T. J., Cheadle, C., and Engel, J. (2003). PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics *4*, 61.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res *32 Database issue*, D23-26.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S.*, et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242-2246.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J.*, et al.* (2004). Ensembl 2004. Nucleic Acids Res *32 Database issue*, D468-470.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T.*, et al.* (2004). An overview of Ensembl. Genome Res *14*, 925-928.

Black, D. L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell *103*, 367-370.

Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. Paper presented at: International Conference on Intelligent Systems for Molecular Biology (Heidelberg).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I.*, et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res *31*, 365-370.

Boue, S., Letunic, I., and Bork, P. (2003). Alternative splicing and evolution. Bioessays *25*, 1031-1034.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G.*, et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

Chiang, J. H., Yu, H. C., and Hsu, H. J. (2004). GIS: a biomedical text-mining system for gene information discovery. Bioinformatics *20*, 120-121.

Cohen, A. M., and Hersh, W. R. (2005). A survey of current work in biomedical text mining. Brief Bioinform *6*, 57-71.

Collier, N., Nobata, C., and Tsujii, J. (2002). Automatic Acquisition and Classification of Terminology using a Tagged Corpus in the Molecular Biology Domain. Terminology *7*, 239-257.

Corney, D. P., Buxton, B. F., Langdon, W. B., and Jones, D. T. (2004). BioRAT: extracting biological information from full-length papers. Bioinformatics *20*, 3206-3213.

Costa, R. M., Yang, T., Huynh, D. P., Pulst, S. M., Viskochil, D. H., Silva, A. J., and Brannan, C. I. (2001). Learning deficits, but normal development and tumor predisposition, in mice lacking exon 23a of Nf1. Nat Genet *27*, 399-405.

Craven, M., and Kumlien, J. (1999). Constructing biological knowledgebases by extracting information from text sources. Paper presented at: AAAI conference on intelligent systems for molecular biology.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics *20*, 604-611.

de Bruijn, B., and Martin, J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inf *67*, 7-18.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K.*, et al.* (2003). PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics *4*, 11.

Dumais, S. T., Platt, D., Hackerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. Paper presented at: ACM-CIKM98.

Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res *25*, 2547-2561.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol *6*, R44.

Ettinger, M. (2002). The complexity of comparing reaction systems. Bioinformatics *18*, 465-469.
Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Towards information extraction: Identifying protein names from biological papers. Paper presented at: 3rd Pacific Symoisium of Biocomputing.

Garcia-Blanco, M. A., Baraniak, A. P., and Lasda, E. L. (2004). Alternative splicing in disease and therapy. Nature biotechnology *22*, 535-546.

Gildea, D., and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. Computational Linguistics *28*, 245-288.

Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., and De Moor, B. (2004). TXTGate: profiling gene groups with text-based information. Genome Biol *5*, R43.

Grabowski, P. J., and Black, D. L. (2001). Alternative RNA splicing in the nervous system. Prog Neurobiol *65*, 289-308.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet *17*, 100-107.

Graveley, B. R. (2002). Sex, AGility, and the regulation of alternative splicing. Cell *109*, 409-412.

Hajic, J., Cmejrek, M., Dorr, B., Ding, Y., Eisner, J., Gildea, D., Koo, T., Parton, K., Penn, G., Redev, D., and Rambow, O. (2004) Natural Language Generation in the Context of Machine Translation, Final Report, The Center for Language and Speech Processing, The Johns Hopkins University.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res *33*, D514-517.

Han, C., Lavoie, B., Palmer, M., Rambow, O., Kittredge, R., Korelsky, T., Kim, N., and Kim, M. (2000). Handling Structural Divergences and Recovering Deropped Arguments in a Korean/English Machine Translation System. Paper presented at: Association for Machine Translation in the Americas 2000 (New York).

Hobbs, J. R., Appelt, D., Israel, D., Bear, J., Kameyama, M., Stickel, M., and Tyson, M. (1997). Fastus: A cascade finite-state transducer for extracting information from natural-language text. In Finite State Devices for Natural Language Processsing, E. Roche, and Y. Schabes, eds. (MIT Press), pp. 383-406.

Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C., and Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE *2005*, pe21.

Joachims, T. (2001). Learning to classify text using support vector machines: methods, theory and algorithms, Kluwer Academic Publishers).

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res *14*, 331-342.

Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics *19 Suppl 1*, i180-182.

Kingsbury, P., and Palmer, M. (2002). From Treebank to PropBank. Paper presented at: 3rd International Conference on Language Resources and Evaluation (LREC-2002) (Las Palmas).

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. Paper presented at: Human Language Technology Conference (San Diego, CA, USA).

Kipper, K., Dang, H. T., and Palmer, M. (2000). Class based construction of a verb lexicon. Paper presented at: 17th National Conference on Artificial Intelligence (AAAI-2000) (Austin, TX).

Kornblihtt, A. R. (2005). Promoter usage and alternative splicing. Curr Opin Cell Biol *17*, 262-268.

Krallinger, M., Erhardt, R. A., and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. Drug Discov Today *10*, 439-445.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet *19*, 640-648.

Lee, C. (2003). Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics *19*, 999-1008.

Leek, T. R. (1997) Information extraction using hidden markov models., University of California, San Diego., San Diego.

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation, University of Chicago Press).

Lewis, S. E. (2005). Gene Ontology: looking backwards and forwards. Genome Biol *6*, 103.

Marcotte, E., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. Bioinformatics *17*, 359-363.

Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization, MIT Press).

Marcus, M. (1994). The Penn Treebank: A revised corpus design for extracting predicate-argument structure. Paper presented at: ARPA Human Language Technology Workshop (Princeton, NJ).

McCallum, A., and Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification.

Meyers, A., Macleod, C., and Grishman, R. (1994). Standardization of the Complement Adjunct Distinction. In Proteus Project Memorandum 64 (New York University, Computer Science Department).

Mika, S., and Rost, B. (2004). Protein names precisely peeled off free text. Bioinformatics *20 Suppl 1*, I241-I247.

Miller, G. A. (1990). WordNet: An on-line lexical database. International Journal of Lexicography *3*, 235-312.

Mirnics, K., and Pevsner, J. (2004). Progress in the use of microarray technology to study the neurobiology of disease. Nat Neurosci *7*, 434-439.

Mitchell, T. M. (1997). Machine Learning, McgrawHill).

Mizuta, Y., and Collier, N. (2004). Zone Indentification in Biology Articles as a Basis for Information Extraction. Paper presented at: Joint Workshop on Natural Language Processing in Biomedicine and its Applications (Geneva, Switzerland).

Muller, H. M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol *2*, e309.

Nello Cristiani, J. S. T. (2000). An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press).

Nelson, S. J., Schopen, M., Schulman, J., and Arluk, N. (2000). An Interlingual Database of MeSH Translations. Paper presented at: 8th International Conference on Medical Librarianship (London, UK).

Nenadic, G., Spasic, I., and Ananiadou, S. (2003). Terminology-driven mining of biomedical literature. Bioinformatics *19*, 938-943.

Netzel, R., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). The way we write. EMBO Rep *4*, 446-451.

Novichkova, S., Egorov, S., and Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics *19*, 1699-1706.

Novichkova, S., Egorov, S., and Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics *19*, 1699-1706.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics *17*, 155-161.

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2001). XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem Sci *26*, 573-575.

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. Nat Genet *31*, 316-319.

Perez-Iratxeta, C., Keer, H. S., Bork, P., and Andrade, M. A. (2002). Computing fuzzy associations for the analysis of biological literature. Biotechniques *32*, 1380-1385.

Pisarra, P., Lupetti, R., Palumbo, A., Napolitano, A., Prota, G., Parmiani, G., Anichini, A., and Sensi, M. (2000). Human melanocytes and melanomas express novel mRNA isoforms of the tyrosinase-related protein-2/DOPAchrome tautomerase gene: molecular and functional characterization. J Invest Dermatol *115*, 48-56.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., and Jurafsky, D. (2004). Shallow Semantic Parsing Using Support Vector Machines. Paper presented at: NAACL-HLT.

Pruitt, K. D., and Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res *29*, 137-140.

Pustejovsky, J., Castano, J., Zhang, J., Kotecki, M., and Cochran, B. (2002). Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. Paper presented at: Pacific Symposium on Biocomputing.

Raychaudhuri, S., Schutze, H., and Altman, R. B. (2002). Using text analysis to identify functionally coherent gene groups. Genome Res *12*, 1582-1590.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics *19*, 368-375.

Ribeiro-Neto, R. B.-Y. a. B. (1999). Modern Information Retrieval, Addison Wesley).

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. Paper presented at: 8th National Conference on Artificial Intelligence (AAAI-93) (The AAAI Press/MIT).

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. Paper presented at: 13th National Conference on Artificial Intelligence (AAAI-96) (The AAAI Press/MIT).

Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting Molecular Binding Relationships from Biomedical Text. Paper presented at: 6th Conference on Applied Natural Language Processing (ANLP-NAACL'2000) (WA).

Roberts, R. J. (2001). PubMed Central: The GenBank of the published literature. Proc Natl Acad Sci U S A *98*, 381-382.

Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. (2004). Extracting Regulatory Gene Expression Networks from PubMed. Paper presented at: 42nd meeting of the Association of Computational Linguistics (Barcelona, Spain).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Paper presented at: International conference on new methods in language processing.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. Bioinformatics *20*, 2597-2604.

Sekimizu, T., Park, H. S., and Tsujii, J. (1998). Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts. Genome Informatics, 62-71.

Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics *4*, 20.

Shah, P.K., Jensen, L.J., Boue, S. and Bork, P. (2005). Extracting transcript diversity from scientific literature. *PLoS Comp Biol.* 1(1): e10

Shatkay, H., and Feldman, R. (2003). Mining the biomedical literature in the genomic era: an overview. J Comput Biol *10*, 821-855.

Srinivasan, P., and Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics *20 Suppl 1*, I290-I296.

St. Laurent, S. (2000). XML Elements of Style (New York, McGraw-Hill).

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using Predicate-Argument Structures for Information Extraction. Paper presented at: 41th Annual Meeting of the Association for Computational Linguistics (Tokyo).

Tanabe, L., and Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. Bioinformatics *18*, 1124-1132.

Tapanainen, P., and Jarvinen, T. (1997). A non-projective dependency parser. Paper presented at: 5th Conference on Applied Natural Language Processing (ANLP'97) (Washington, D.C.).

Tateisi, Y., Ohta, T., and Tsujii, J. (2004). Annotation of Predicate-argument Structure on Molecular Biology Text. Paper presented at: Workshop on the 1st International Joint Conference on Natural Language Processing (IJCNLP-04) (China).

Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J. J., Le Texier, V., and Muilu, J. (2004). ASD: the Alternative Splicing Database. Nucleic Acids Res *32 Database issue*, D64-69.

Vapnik, V. N. (1999). The nature of statistical learning theory, 2 edn, Springer).

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acids Res *31*, 258-261.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P*., et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wattarujeekrit, T., Shah, P., and Collier, N. (2004). PASBio:predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics *5*, 155.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E*., et al.* (2004). Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res *32 Database issue*, D35-40.

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res *30*, 3754-3766.
Yan, J., and Marr, T. G. (2005). Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res *15*, 369-375.

Yangarber, R. (2003). Counter-Training in Discovery of Semantic Patterns. Paper presented at: 41st Annual Meeting of the Association for Computational Linguistics (Tokyo).

Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000). Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. Paper presented at: 6th Conference on Applied Natural Language Processing (ANLP-NAACL'2000) (WA).

Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004). Variation in alternative splicing across human tissues. Genome Biol *5*, R74.

Yiming Yang, X. L. (1999). A re-examination of text categorization methods. Proceedings of ACM SIGIR Conference on Research and Developement in Information Retrieval, 42-49.

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D. A., Hayashizaki, Y., and Gaasterland, T. (2003). Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res *13*, 1290-1300.

Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. Bioinformatics *20*, 1178-1190.