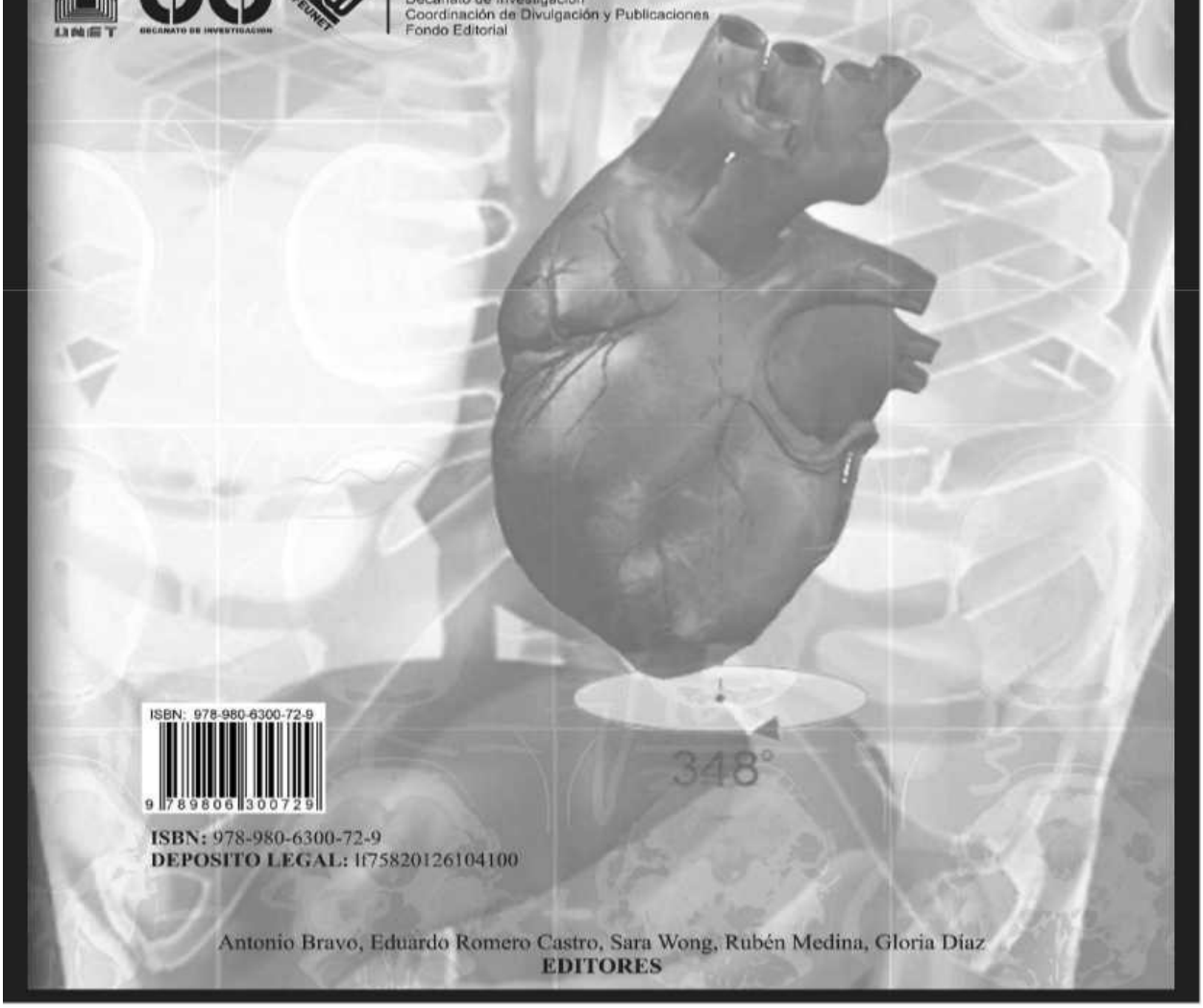
VIII INTERNATIONAL SEMINAR ON MEDICAL INFORMATION PROCESSING AND ANALYSIS

Universidad Nacional Experimental del Táchira
Vicerrectorado Académico
Decanato de Investigación
Coordinación de Divulgación y Publicaciones
Fondo Editorial



ISBN: 978-980-6300-72-9

ISBN: 978-980-6300-72-9
DEPOSITO LEGAL: lf75820126104100

Antonio Bravo, Eduardo Romero Castro, Sara Wong, Rubén Medina, Gloria Díaz
EDITORES

BioPattern: Text-mining PLN para el análisis automático de eventos regulatorios en redes de señalización biológica.

Arquitectura y avances.

J. López[1], D. Rubio[1], K. Sandoval[1], J. Dávila[2], D. Nimo[1], M. Bernal[1],

J. Maldonado[1], J. Rui[1] .

[1]Laboratorio de Computación de Alto Rendimiento, Universidad del Táchira,

San Cristóbal, Táchira, Venezuela.

jlopez@unet.edu.ve

[2]CESIMO: Centro de Simulación y Modelos, Fac. Ingeniería.

Universidad de Los Andes, Mérida, Venezuela.

jacinto@ula.ve.

Resumen: El proyecto aquí presentado trata el problema del modelado y análisis de redes de señalización biológica. En particular, se presenta una estrategia para identificar señales (motivos) en una secuencia de ADN problema, que podrían intervenir en la regulación del proceso de transcripción correspondiente. Para tal fin se presenta un sistema al que se denomina BioPattern, organizado en dos componentes generales. El primer componente, llamado BM, define un buscador de motivos que explora y coteja, contra bases de datos especializadas, la presencia de motivos candidatos en la secuencia problema. El segundo componente, llamado Rpln, emplea las salidas del componente BM para realizar una forma de *text-mining* de publicaciones científicas, en la búsqueda de descripciones de eventos regulatorios expresables como triadas (o tripletas) lógicas del tipo (sujeto, relación, objeto). El desarrollo de experimentos para un conjunto de objetos biológicos definido por parte de expertos, establece que BioPattern puede emplearse en la construcción automática o semiautomática de redes de señalización. El sistema logra la detección de eventos contenidos en un mapa de señalización, lo que sugiere que no solo podría usarse para reconstruir tal mapa, si no además para incorporar conocimiento aún no considerado por la vía manual. Se presenta el modo en que los procesos mencionados se llevan a cabo y el tipo de resultados obtenidos.

Palabras clave: Text-Mining, Redes de Señalización Biológica, Bioinformática, Genética.

1 Introducción

El proyecto aquí presentado trata el problema del modelado y análisis de redes de señalización biológica [1]. En particular, se presenta una estrategia para identificar señales, en una secuencia de ADN problema, que intervienen en la regulación del proceso de transcripción correspondiente. La identificación de esas señales puede establecer modos alternativos de transcripción que podrían ser claves para el diseño de fármacos o identificar el sitio de inicio del posible gen portado por la secuencia en estudio. Diversos recursos informáticos y estrategias computacionales han sido desarrollados a tal fin [2-11].

El problema planteado requiere el uso de diversos procesos computacionales: búsquedas de genes evolutivamente afines a la secuencia problema (consulta de homólogos [12]), estudios comparativos entre secuencias para definir motivos candidatos (uso de alineamientos [12]), validación de motivos candidatos mediante información disponible en la Internet (uso de Servicios Web [13]); el análisis automatizado de publicaciones, que aporten información científica relativa a los motivos candidatos y modos de interacción de estos con objetos biológicos diversos (uso de Servicios Web y Procesamiento de Lenguaje Natural [14]) y, por último, el desarrollo de procesos de inferencia que deduzcan los posibles patrones de señalización, asociables a una secuencia problema [15-16].

BioPattern se ha organizado en dos componentes generales (ver Fig. 1). El primer componente define un buscador de motivos que explora el alineamiento de homólogos y coteja, contra bases de datos especializadas, la presencia de motivos candidatos en la secuencia problema. No cualquier combinación de motivos candidatos participa en un patrón de señalización. Las publicaciones científicas pueden explicar cuáles si y cuáles no. Para recuperar e incorporar ese conocimiento, se usa el segundo componente del sistema. Este componente realiza una forma de *text-mining* [17] de publicaciones científicas en la búsqueda de descripciones de eventos regulatorios expresables como triadas (o tripletas) lógicas (sujeto, relación, objeto). Sujetos y objetos de una triada corresponden a objetos biológicos como ligandos, proteínas, genes y motivos, asociados a la secuencia en estudio. Son ejemplos de “relación” aquellas que tienen que ver con la estimulación, inhibición, activación o facilitación, del proceso de transcripción del posible gen portado por la secuencia problema.

El componente explorador de triadas, al que llamaremos Rpln, emplea una versión del resumidor de textos basado en PLN, descrito en [18-19]. El resumidor recibe una publicación y devuelve un compendio de tríadas describiendo relaciones relevantes, extraídas de ese texto. Para precisar las relaciones, el resumidor emplea un diccionario de verbos especializado en el tema del que tratan las publicaciones, en este caso, redes de señalización biológica. Se presentan ejemplos del modo en que tales procesos se llevan a cabo y del tipo de resultados correspondientes.

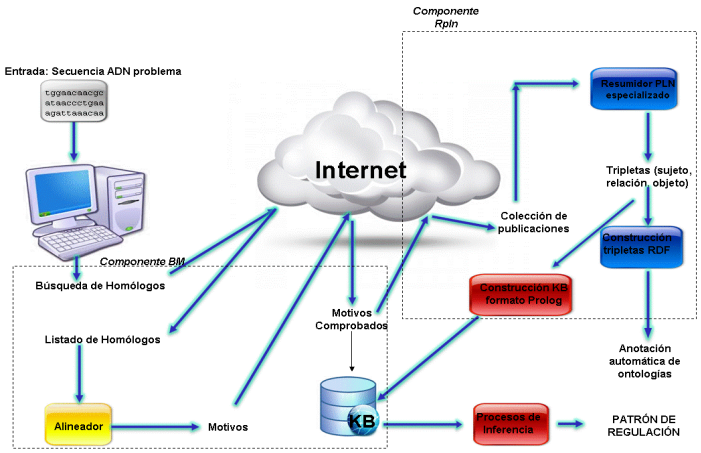


Fig. 1. Arquitectura para la sistematización de la búsqueda de patrones de señalización.

2 La Estrategia y sus Componentes

La Fig. 1 ilustra los componentes generales del sistema: Buscador de Motivos (BM) y el Resumidor de textos basado en PLN (Rpln). El BM propone posibles señales reguladoras para una secuencia de ADN problema. Dada la secuencia, el BM recupera otras secuencias ADN homólogas, vía un servicio Web como el disponible en el NCBI (*The National Center for Biotechnology Information*) [11] y descrito en *Entrez Programming Utilities* Help [20]. Las homólogas así obtenidas son empleadas para definir sus regiones promotoras, también mediante consultas a los servicios del NCBI. Tales regiones son alineadas con la región promotora de la secuencia problema, por medio de ClustalW [12], lo que define bloques comunes presentes en todas las regiones promotoras. Posteriormente, los motivos candidatos (señales reguladoras) para la secuencia problema, son obtenidos explorando en los bloques coincidentes hallados. Esto último es realizado

por BM empleando consultas a un servicio Web, especializado en el reporte de motivos experimentalmente comprobados.

EL tipo de resultados obtenidos con el componente BM puede verse en la Tabla 1. En la Tabla 1 se muestran: la secuencia problema, el grupo de secuencias homólogas que se alinean con la secuencia problema y un conjunto de motivos de referencia. La salida del componente BM consiste del alineamiento (consenso) de las secuencias y el conjunto de motivos comprobados. Los motivos validados por el componente BM son empleados como palabras claves para la selección de publicaciones, en las que se reporte su presencia como elementos reguladores. Las publicaciones obtenidas son procesadas por el componente Rpln, a fin de construir una base de conocimiento de hechos regulatorios del tipo (sujeto, relación, objeto).

El componente Rpln es un módulo de extracción de información basado en el resumidor de textos descrito en [18-19]. El resumidor ha sido extendido con una interfaz a una ontología (ver Fig. 2) y respecto a ella debe proceder para extraer un listado de posibles relaciones-acciones (verbos y sus conjugaciones), particulares al área de estudio. En este trabajo la ontología suministrada corresponde al sistema biológico descrito en [21] [22]. El componente Rpln produce oraciones etiquetadas con la estructura ([sujeto(X), verbo(Y), complemento(Z)]). En esta estructura, el complemento suele contener el objeto afectado por el sujeto, a través de la acción descrita por el verbo; lo que permite construir las tríadas del tipo (sujeto, relación, objeto). De esta forma, la aplicación del componente Rpln a una colección de publicaciones, produce una base de conocimiento sobre eventos de regulación. Las Tablas 2 muestran un ejemplo de la aplicación del componente Rpln, resaltando un ejemplo a partir de la cual puede construirse el evento regulatorio (*p53*, *ejerce-expresión*, *gadd45*).

3 Experimentos iniciales

Para validar el sistema integrado BM+Rpln se asumen que: *Dada una red biológica de señalización construida por expertos, el sistema debería ser capaz de detectar eventos presentes en tal red, dado que se le suministren términos relativos a objetos biológicos inherentes a la misma*. Considérese la parte superior de la Fig. 3, correspondiente a una región extraída del mapa reportado en [21][22]. Allí puede observarse la enzima CYP1A1, además de otras enzimas cercanas a la misma (CYP1A2, CYP1B1). Al respecto se describe un experimento tipo.

El primer paso consiste en emplear el sistema Ensembl [23] para descargar la secuencia de ADN, correspondiente al gen que codifica la enzima CYP1A1. A partir de allí se procede a la búsqueda de genes homólogos, haciendo uso del sistema HomoloGen [24],

disponible en NCBI [11]. Definidas las secuencias homólogas se procede a definir regiones promotoras para cada secuencia. Esto último se realizó siguiendo los criterios descritos en [25]. Realizado lo anterior se activa el componente BM de BioPattern, suministrando las secuencias y las regiones promotoras previamente organizadas. En este punto CYP1A1 es una proteína, para la que se desea hallar posibles eventos regulatorios. BioPattern propone motivos candidatos, posiblemente involucrados en la señalización de la transcripción de la secuencia correspondiente a CYP1A1. Seguidamente, se procede a la búsqueda de publicaciones, que pudieran reportar eventos y objetos biológicos afines a la secuencia. Los resultados (parciales) correspondientes pueden verse en la parte inferior de la Fig. 3.

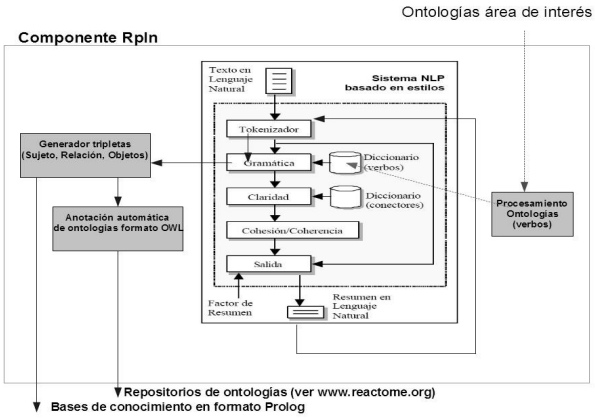


Fig. 2. Componente Rpln

Tabla 1. Salida componente BM para una secuencia problema.

4 Resultados

Los resultados obtenidos pueden ser analizados observando la Fig. 3 en su parte baja. Para ello debe observarse que existen relaciones descritas en la parte superior de la misma figura, que logran ser reportadas independientemente por el sistema. Tómese por ejemplo la primera oración. Allí se indica que "*similarmente a como ocurre a CYP1a1 y CYP1b1, la transcripción de CYP1a2 es principalmente regulada por el receptor aromático de hidrocarbono (ahr), que resulta ser un receptor activado por ligando*". El evento regulatorio recién descrito puede ser identificado en la parte alta de la Fig. 3. Allí se puede observar que el receptor *ahr* interactúa con el receptor *arnt*, conformando un complejo (representado por el punto que emerge de la línea que los une). El complejo conformado interactúa a su vez con el motivo XRE, ubicado en la región promotora del gen codificante de la proteína *CYP1A2*, lo que finalmente activa su transcripción. Nótese que en esa misma región del mapa se representan líneas que llegan a las proteínas *CYP1A1* y *CYP1B2*, lo que indica que las tres relaciones descritas en la oración, aparecen representadas en el mapa. Experimentos similares se desarrollan para otros objetos del mapa, lo que conduce a su representación en una base de conocimiento, al estilo de la presentada en la parte derecha de la figura 3.

[[sujeto([la, concentración, celular, de, p53]), [verbo([debe]), verbo([estar])], complemento([fuertemente, regulada, (,), ya, que, aunque, puede, suprimir, tumores, (,), el, alto, nivel, de, p53, puede, acelerar, el, proceso, del, envejecimiento, por, apoptosis, excesiva])], [sujeto([por, esta, razón, en, células, sanas, p53]),

verbo([tiene]), complemento([una, vida, media, corta, (, 20, min,)])], [sujeto([cuando]), [sujeto_verbo([se]), verbo([produce])], complemento([un, ataque, a, la, célula, y, se, produce, daño, en, el, DNA, (,), p53, detecta, la, presencia, de, daño, celular])], [sujeto([los, dos, sensores, fundamentales, del, nado, en, el, DNA]), verbo([son]), complemento([dos, kinasas, relacionadas, :, atm, (, por, ataxia, telangiectasia, mutated,), y, atr, (, por, ataxia, telangiectasia, and, rad3, related,)])], [sujeto([la, fosforilación, de, p53, la, libera, de, su, asociación, con, mdm2, (,), por, lo, que, su, vida, media, aumenta, y]), [verbo([puede]), verbo([ejercer])], complemento([su, función, de, factor, transcripcional, (,), aumentando, la, expresión, de, genes, importantes, para, la, reparación, del, daño, en, el, DNA, (, como, gadd45,), (,), para, inhibir, la, progresión, a, través, del, ciclo, celular, (, como, p21,), y, para, promover, la, apoptosis, en, caso, necesario, (, como, bax,)])]]

Tabla 2. Resumen obtenido para el texto indicado en la Tabla 2.

5 Conclusiones

El resultado, ya sea que se trate de una secuencia desconocida o no, es una base de conocimiento que describe objetos y eventos regulatorios con ella relacionados. El desarrollo de experimentos para un conjunto de objetos biológicos bien definido por parte de expertos [21][22], establece que BioPattern puede emplearse en la construcción automática o semiautomática de redes de señalización. El sistema logra la detección de eventos contenidos en el mapa del sistema BAXS [21][22], lo que sugiere que no solo podría reconstruirse el mapa allí descrito, si no además incorporar conocimiento aún no considerado por la vía manual.

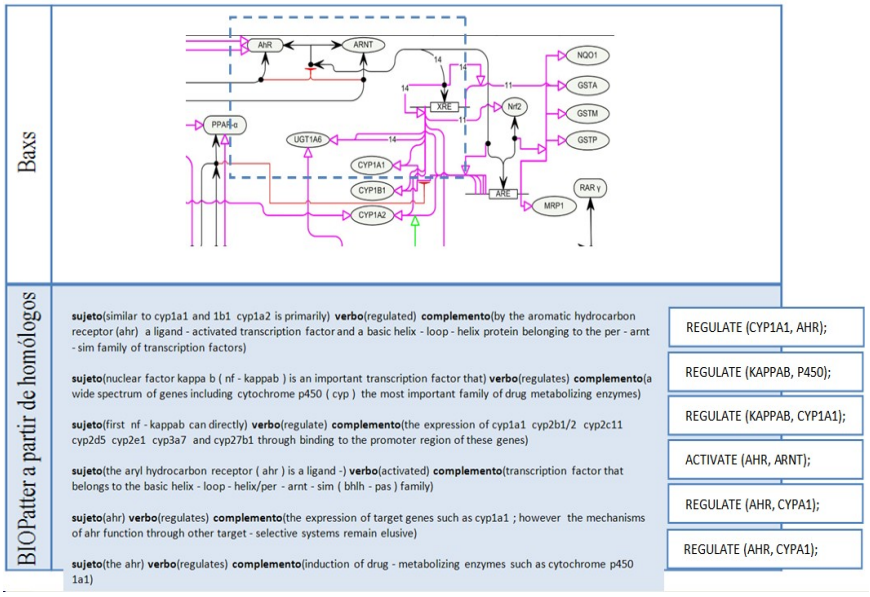


Fig. 3. Resultados para la enzima CYP1A1 a partir de homólogos extraídos del mapa del sistema BAXS [22].

6 Trabajo Futuro

Debe medirse la efectividad del sistema implementado comparando los eventos regulatorios minados por BioPattern y aquellas presentes en el mapa del sistema BAXS. Tales experimentos permitirán la generación de una nueva versión del mapa susceptible de análisis tanto humano como artificial. Debe tratarse la sistematización de las consultas a bases de datos externas, tanto de motivos y genes homólogos, como las relativas a publicaciones. Para ello se están estudiando recursos ofrecidos por centros especializados en bioinformática, como el NCBI [11] y el EBI [9]. El componente analítico del sistema requiere la definición de métodos para explorar las bases de conocimiento. Esto último puede ser abordado mediante distintas estrategias de análisis inferencial [15][16].

Referencias

1. Davidson E, Levin M. "Gene regulatory networks". Proc. Natl. Acad. Sci. U.S.A. 102 (14): 4935. DOI:10.1073/pnas.0502024102. PMC 556010. PMID 15809445. (2005).
2. Bentley DR. The Human Genome Project: an overview (2000). Med Res Rev. May;20(3):189-96.
3. Salzberg S.L., D.B. Searls, S. Kasif (1998), Computational Methods in Molecular Biology, Elsevier.
4. Michael P Cary, Gary D Bader, and Chris Sander. Pathway information for systems biology. FEBS Lett, 579(8):1815–1820, Mar 2005.
5. Burge C, Karlin S: Prediction of complete gene structures in human genomic ADN. J Mol Biol 1997, 268:78-94
6. The GENSCAN Web Server at MIT. http:// http://genes.mit.edu/GENSCAN.html.
7. GenBank. http://www.ncbi.nlm.nih.gov/genbank/
8. PubMed Home Page. http://www.pubmed.org.
9. European Bioinformatics Institute Home Page. http://www.ebi.a.uk.
10. Ensembl Genome Browser. http://www.ensembl.org/index.html.
11. NCBI (The National Center for Biotechnology Information). http://www.ncbi.nlm.nih.gov/
12. Thompson JD , Gibson TJ , Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics (ISSN: 1934-3396). 2002 Aug; Volume: Chapter 2.
13. Web of Services. http://www.w3.org/standards/webofservices/
14. Jurafsky, D., Martin., J.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2009.
15. Muggleton S.H. (2005); Machine learning for systems biology. In Proceedings of the 15th International Conference on Inductive Logic Programming, LNAI 3625, pages 416-423. Springer-Verlag, 2005.
16. Baral, C., K. Chancellor, N. Tran, N. L. Tran, A. Joy, and M. Berens (2004). A knowledge based approach for representing and reasoning about signalling networks. Bioinformatics, 20 Suppl 1:I15–I22.

17. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press. 2008. ISBN 0521836573.
18. Dávila Jacinto y H. Yelitza Contreras (2002). Una gramática de estilos para resumir textos en español. Revista de La Sociedad Español para el Procesamiento del Lenguaje Natural (SEPLN). Valladolid 11 al 13 de Septiembre de 2002. (Indizada en Latindex).
19. Parra María Marilú y Jacinto Dávila (2005). Un Modelo Computacional para la Generacion de Resumenes Automaticos de Articulos Cientificos en Español. ACTAS del IX Simposio Internacional de Comunicación Social. Santiago de Cuba, 24 al 28 de Enero, CUBA, 2005.
20. Entrez Programming Utilities Help. Bethesda (MD) (2010): National Center for Biotechnology Information (US). http://www.ncbi.nlm.nih.gov/books/NBK25501/
21. Schmidt Oliver, Jose Lopez , Francisco Azuaje, Paul Thompson, Werner Dubitzky (2009); A standard-compliant global map of the bile acid/xenobiotic signalling network: Construction and automated query processing; The 2009 International Conference on Bioinformatics and Computational Biology (BIOCOMP'2009).
22. BAXS Web Reasoner. http://baxs.scic.ulst.ac.uk:8080/BAXSWebReasoner/.
23. Ensembl Genome Browser. http://www.ensembl.org/index.html.
24. HomoloGene. http://www.ncbi.nlm.nih.gov/homologene.
25. Mary C. Thomas and Cheng-Ming Chiang. “The General Transcription Machinery and General Cofactors”. Department of Biochemistry. Critical Reviews in Biochemistry and Molecular Biology, 41:105–178, 2006. Case Western Reserve. University School of Medicine. Cleveland, OH, USA.