Título: TECNOLOGÍA DIFUSA PARA GENE ONTOLOGY Y SITIOS DE UNIÓN DE FACTORES DE TRANSCRIPCIÓN

Nombre: García Alcalde, Fernando

Universidad: Universidad de Granada

Departamento: CIENCIAS DE LA COMPUTACION E INTELIGENCIA ARITIFICIAL

Fecha de lectura: 18/02/2010

Mención a doctor europeo: concedido

Programa de doctorado: DISEÑO, ANÁLISIS Y APLICACIONES DE SISTEMAS INTELIGENTES

Dirección:

> **Director:** Armando Blanco Morón

Tribunal:

> **presidente:** ANTONIA ARANEGA JIMENEZ

> **secretario:** David Pelta

> **vocal:** ANTONIO MARIN RODRIGUEZ

> **vocal:** Joaquín Dopazo Blázquez

> **vocal:** Jack Leunissen

Descriptores:

> INFORMATICA

> BIOLOGIA CELULAR

El fichero de tesis ya ha sido incorporado al sistema

> https://digibug.ugr.es/flexpaper/handle/10481/4862/18622045.pdf?sequence=1&isAllowed=y

Resumen: Motivación:
En la última década, la bioinformática se ha convertido en una parte integral de la investigación y el desarrollo en las ciencias biomédicas. La bioinformática tiene ahora un papel esencial en el desciframiento de datos genómicos, transcriptómicos y proteómicos generados por tecnologías experimentales de alto rendimiento, y en la organización de la información obtenida por la biología tradicional. Los métodos de análisis de secuencias de genes o proteínas han evolucionado y mejorado, desarrollándose nuevos métodos para el análisis de un gran número de genes o proteínas simultáneamente, así como para la identificación de grupos de genes relacionados y redes de interacción de proteínas. Conseguida la secuenciación de los genomas de un número cada vez más alto de organismos, la bioinformática está comenzando a ofrecer tanto las bases conceptuales como los métodos prácticos para la detección de

conductas funcionales sistémicas de la célula y el organismo.
La bioinformática es, por tanto, el campo de la ciencia donde la biología, la informática, y la tecnología de la información se unen para formar una única disciplina, con el objetivo de ayudar en el descubrimiento de nuevos datos biológicos. De esta forma, la comprensión de los principios biológicos que afectan a los organismos vivos es clave para el desarrollo de métodos bioinformáticos apropiados.
A lo largo de la historia de la ciencia, siempre ha existido la necesidad de modelar y gestionar la incertidumbre existente en los experimentos reales. Esto es particularmente cierto en la biología en general, y más recientemente en la bioinformática. La variabilidad exhibida por la naturaleza al estudiar el genoma y sus relaciones requieren modelos computacionales lo suficientemente flexibles para capturar lo esencial, sin tener en cuenta todas las variabilidades como algo completamente nuevo. La teoría difusa (Zadeh, 1965) es una potente herramienta que ha servido a los investigadores para el modelo de situaciones donde la principal fuente de incertidumbre es la aleatoriedad.
En algunos casos, la incertidumbre puede adoptar otras formas: Al considerar una secuencia nueva de un gen, puede ser de interés conocer cómo de similar es a otra secuencia en particular. No se trata de el clásico problema binario de saber si dos secuencias son iguales o no, si no de saber cuánto se asemejan estas dos secuencias. Otras fuentes de incertidumbre incluyen: omisiones en los datos extraídos de muestras reales, la falta de expresividad o de confianza en algunas características extraídas, la falta de límites claros entre las distintas clases de proteínas, genes o productos de los genes que son miembros de más de una clase, etc.
Además, hasta la fecha casi todos los problemas bioinformáticos se han formulado de un modo determinista. La mayoría de estos problemas son definidos fijando y optimizando funciones objetivo. Sin embargo, existen diversas situaciones en las que se hace necesario considerar la vaguedad de los datos. Por ejemplo, la imprecisión que acompaña intrínsecamente a los sistemas biológicos, las múltiples funciones que una entidad biológica puede desarrollar, las descripciones difusas de algunos fenómenos biológicos, etc.
Esta tesis tiene como objetivo principal resolver algunos importantes problemas bioinformáticos mediante la aplicación de nociones sobre la teoría de conjuntos difusos y lógica difusa, así como sobre otros métodos de soft computing.
Debido a la reciente explosión de nuevos datos biológicos procedentes de la secuenciación de genomas, los científicos se enfrentan al problema de responder a muchas preguntas básicas, tratando de extraer información de estos datos. Una de las principales tareas es la de descubrir la función a la que los genes están asociados. Recientemente, la bio-ontologías han desempeñado un papel importante para la integración automática de conocimiento, lo que es fundamental para apoyar la generación y validación de hipótesis sobre las funciones de los productos de los genes. En este sentido, la Gene

Ontology1 (GO) (Ashburner etal., 2000) se ha convertido en un estándar de facto para describir los productos de genes. Fue creada con el objetivo de normalizar la representación de los genes y productos de genes procedentes de distintas especies y bases de datos. Así, proporciona un vocabulario estructurado y controlado para describir las funciones de los genes y los productos de genes en cualquier organismo. Existen algunos trabajos que utilizan GO para extraer la información biológica de grupos de productos de genes potencialmente relacionados. Por lo general, estos enfoques se basan en medidas definidas para una ontología genérica que son adaptadas a las características específicas de GO (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998).
Sin embargo, pocos métodos basados en tecnología difusa están disponibles actualmente. En nuestra opinión, las propiedades de la teoría de conjuntos difusos la hacen interesante para su aplicación a este problema.
Por otro lado, las células controlan la abundancia de proteínas por medio de diversos mecanismos. Uno de esos mecanismos es la regulación de la transcripción, que es un proceso continuo en el que muchos factores se combinan para garantizar una adecuada tasa de síntesis de proteínas. La comprensión de estos complejos procesos es uno de los principales objetivos de la biología computacional. Los factores de transcripción (TFs, del inglés transcription factors) desempeñan un papel clave en la regulación de los genes, mediante su unión a secuencias específicas, llamadas sitios de unión de factores de transcripción (TFBSs, del inglés transcription factor binding sites).
Aquellas secuencias de ADN donde se puede producir la unión del mismo TF se agrupan conjuntamente formando motivos, que se representan normalmente como matrices de frecuencias por posición (PFMs, del inglés position frequency matrices). La predicción in silico de la posible unión de un TF a un TFBS es un problema ampliamente estudiado por la biología computacional.
Un punto importante en el contexto del descubrimiento de novo de motivos consiste en saber si, dados los candidatos a motivos recientemente obtenidos, éstos se asemejan a otros motivos previamente descritos en las bases de datos existentes. Por ésta y otras razones se han desarrollado varias medidas de comparación entre motivos. La mayoría de los métodos propuestos están basados en técnicas estadísticas que comprueban si las diferentes columnas de los motivos pertenecen a la misma distribución (Pietrokovski, 1996; Schones etal., 2005; Wang and Stormo, 2003). Otros trabajos más recientes proponen el uso de métodos más específicos que mejoran a los enfoques probabilísticos (Gupta etal., 2007; Pape etal., 2008). Sin embargo, en el contexto de la comparaciones entre motivos, la utilización de PFMs como representación de las preferencias de unión de los TFs incluye imprecisión. Además, los métodos actuales no están diseñados para tener en cuenta la mayor aportación de las posiciones mejor conservadas de los motivos a la fuerza de la unión secuencia-motivo. Por lo tanto, nuevos métodos que tengan en cuenta este tipo de problemas son necesarios.

Del mismo modo, el descubrimiento de patrones en secuencias de ADN es una de los problemas más importante de la bioinformática, con aplicaciones en la búsqueda de elementos de regulación y TFBSs. Una importante tarea en este problema es la búsqueda (o predicción) de sitios de unión conocidos en una nueva secuencia de ADN. La mayoría de las herramientas disponibles para la predicción de TFBSs asumen independencia entre las posiciones de las bases de los sitios de unión (Hertz etal., 1990; Sandelin etal., 2004b). Algunos trabajos recientes están empezando a considerar las dependencias entre posiciones (Tomovic and Oakeley, 2007; Zare-Mirakabad etal., 2009). Uno de los objetivos principales en la predicción de TFBSs es reducir la tasa de falsos positivos sin comprometer la sensibilidad de los resultados. Los métodos que tienen en cuenta las dependencias entre posiciones tienden a ser significativamente más eficaces. Sin embargo, algunas cuestiones como el sobreaprendizaje de las secuencias de prueba, o la selección de un umbral arbitrario para detectar las dependencias entre posiciones, siguen abiertas y necesitan nuevos enfoques para su resolución.

Objetivos

El objetivo general de esta tesis es encontrar soluciones basadas en la tecnología difusa para algunos problemas bioinformáticos importantes, gestionando así la incertidumbre asociada a los procesos biológicos. Más concretamente, nos centramos en el estudio de las medidas de similitud semántica para GO, las comparaciones de motivos de ADN, y la cuantificación de la afinidad secuencia-motivo.

De acuerdo a esto, los objetivos específicos de esta tesis son los siguientes:

* Analizar las propiedades de GO y revisar el estado del arte de las medidas semánticas descritas sobre GO.
* Comparar las medidas semánticas crisp2 sobre GO y analizar sus limitaciones.
* Aplicar diferentes métodos de agrupamiento y comparar su utilidad para el reconocimiento de familias de proteínas.
* Proponer una nueva medida de similitud semántica difusa para GO.
* Incorporar a la medida los códigos de evidencia de las anotaciones para tener en cuenta la fiabilidad de la fuente de información.
* Comparar la nueva medida con las medidas existentes en problemas de clasificación de proteínas.
* Revisar el estado del arte de las medidas de comparación entre motivos y examinar la adecuación de enfoques difusos para dicha tarea.
* Adaptar medidas difusas clásicas al problema de la comparación de motivos.
* Comparar las medidas difusas con otros enfoques relacionados en problemas de detección de motivos.

* Proponer una nueva medida de similitud entre motivos basada en la integral difusa, diseñada teniendo en cuenta la distinta importancia de las posiciones de los motivos en función de su contenido de información
* Revisar medidas entre motivos recientes y analizar sus inconvenientes.
* Definir la nueva medida y demostrar su mejor funcionamiento en experimentos tanto sintéticos como reales.
* Proponer un nuevo método basado en tecnología difusa para cuantificar la afinidad secuencia-motivo.
* Discutir los últimos avances en esta materia.
* Mejorar la calidad de la predicción de TFs de los enfoques existentes.
* Aplicar el nuevo método a problemas biológicos reales.

Contenidos

Esta memoria se estructura en cuatro partes bien diferenciadas, cada una de las cuales se compone de uno o más capítulos.

La Parte I contiene el Capítulo 1 que incluye una introducción en la que, partiendo de los antecedentes en el área, se motiva nuestro trabajo, se establecen los objetivos de la tesis y se describe el contenido del documento.

A continuación, la Parte II expone los conocimientos preliminares necesarios para una mejor comprensión del texto. El Capítulo 2 presenta algunos conceptos básicos de biología, así como proporciona una revisión de la bioinformática, el campo de investigación donde se enmarca esta tesis. El Capítulo 3 introduce algunas nociones básicas sobre la teoría de conjuntos difusos, la lógica difusa y otros métodos de soft computing.

La Parte III presenta las contribuciones de esta tesis. El Capítulo 4 define una nueva medida de similitud difusa para GO, e investiga el funcionamiento de ésta y otras medidas semánticas para GO en conjunto con distintos métodos de agrupamiento en problemas de clasificación de proteínas. El Capítulo 5 expone el problema de comparación entre motivos de TFBSs, y muestra cómo se pueden adaptar para esta tarea distintas medidas difusas clásicas. El Capítulo 6 introduce los avances más recientes de las medidas entre motivos y presenta una nueva medida de similitud para motivos de ADN basada en la integral difusa llamada FISim. Además, propone una nueva metodología de agrupamiento basada en dicha medida y en métodos de kernelización. El capítulo termina con una evaluación de nuestras propuestas en comparación con los mejores métodos existentes. El Capítulo 7 aborda el problema de búsqueda (o predicción) de sitios de unión ya conocidos en una secuencia de ADN, proponiendo una nueva medida de afinidad motivo-secuencia basada en la teoría de conjuntos difusos intuicionista.

Finalmente, la Parte IV, concluye esta tesis. El Capítulo 8 resume las contribuciones de esta tesis. Los resultados se analizan de acuerdo con los objetivos anteriormente establecidos. Además, se apuntan algunas ideas para el

trabajo futuro.