Minería de Datos en Bio-Ciencias

Prato Laura Beatriz[1], Fresno Rodríguez Cristóbal[2], Fernandez Elmer Andres[2], Zingaretti María Laura[1], Ribero Gabriela[1], Villoria Liliana[1]
[1]Instituto AP de Ciencias Básicas y Aplicadas - Universidad Nacional de Villa María
[2]Grupo de Minería de Datos - Universidad Católica de Córdoba

Universidad Nacional de Villa María: Av. Arturo Jauretche 1555 – Villa María – Córdoba – Argentina
0353-4539106 / 141
lprato@unvm.edu.ar - cristobalfresno@gmail.com – elmerfer@gmail.com

Resumen

El campo de las Bio-ciencias está en pleno desarrollo y expansión. La variedad de tecnologías disponibles y aplicaciones están generando cantidades abrumadoras de datos que necesitan de protocolos, conceptos y métodos que permitan un análisis uniforme y asequible.

Otra característica distintiva de estos ámbitos es su condición multidisciplinaria, donde interactúan (y cada vez más) disciplinas como la biología, la matemática, la estadística, la informática, la inteligencia artificial; y sus aplicaciones sobre la agronomía, la salud humana y animal y el medio ambiente; por lo que cualquier esfuerzo tendiente a aumentar el nivel de comunicación y entendimiento entre las distintas disciplinas redundará en beneficios.

La Minería de Datos, concepto que aglutina una variedad de metodologías analíticas, proporciona un marco conceptual y metodológico para el abordaje del análisis de datos y señales en distintas disciplinas. Sin embargo cada campo de aplicación presenta desafíos específicos que deben ser abordados particularmente desde la racionalización de los conceptos específicos del ámbito.

En este proyecto se integrarán las experiencias y criterios de distintas disciplinas que están involucradas en el desarrollo experimental en bio-ciencias. La finalidad es elaborar protocolos y metodologías de análisis, desarrollar métodos analíticos para generar nuevas estrategias diagnósticas, predictivas a partir de los datos recogidos que permitan extraer conocimiento en problemas biotecnológicos que se basen en investigación sólida de los procedimientos estadísticos/bioinformáticos relevante para el manejo de datos experimentales.

Palabras clave:

Bioinformática - minería de datos – biotecnología - inteligencia artificial

Contexto

Esta línea se inserta en un proyecto de desarrollo conjunto entre la Universidad Nacional de Villa María y la Universidad Católica de Córdoba. Ha sido evaluado por evaluadores externos, participa del Programa de Incentivos de la Nación para Docentes Investigadores, y es financiado por la Universidad Nacional de Villa María

Introducción

El dominio de técnicas de análisis de Bio-datos, como la capacidad para diseñar y analizar experimentos cuyos resultados puedan ser luego transformados en nuevos dispositivos o técnicas de diagnóstico, es fundamental para hacer un uso más eficiente de los recursos destinados a la investigación y al desarrollo de productos biotecnológicos (vacunas, dispositivos de diagnóstico biomédico, mejoramiento de especies,

evaluación de drogas, mejoramiento vegetal y animal, calidad de alimentos, etc).

Los pasos involucrados en la adquisición, análisis e interpretación de Bio-datos son numerosos y su correcto abordaje es crucial para el éxito de la aplicación [Buckingham2003]. El modelado y la búsqueda de patrones tendientes al diseño de un método de diagnóstico o caracterización implican la utilización de estrategias analíticas capaces de interrelacionar una gran cantidad de variables con un evento particular. Los métodos de PLS y SVM son métodos modernos y prometedores para el análisis y modelado de señales y datos biomédicos/biotecnológicos, más aun si existen mecanismos que permitan modelar en cierto grado (a través de núcleos adecuados) algún tipo de conocimiento específico (prior-knowledge). La inclusión de conocimiento específico en sistemas de análisis inteligente es un tema candente en investigación de Minería de Datos y Reconocimiento de Patrones. Dominar este tipo de técnicas permitirá evaluar rápidamente su aplicabilidad en distintos ámbitos. Unos de los ámbitos que está demandando este tipo de metodologías de análisis es la genómica y proteómica de alto rendimiento, donde Argentina cuenta actualmente con secuenciadores masivos y está adquiriendo nuevo equipamiento a través de las convocatorias a plataforma PPL dado que están consideradas áreas de vacancia. En este contexto y en otros como la agrobiotecnología, la generación masiva de datos experimentales está requiriendo de la adaptación y/o creación de herramientas específicas para satisfacer los requerimientos de estas tecnologías y poder extraer información biológica relevante. En este proyecto se pretende desarrollar metodologías estadístico/computacionales para proveer de técnicas de minería de datos a estos sectores de investigación y productivos de la región a través de un enfoque multidisciplinar, integrando conceptos y algoritmos provenientes de la Bioinformática, la Estadística, la Bioingeniería, la Biología y la Inteligencia Artificial, disciplinas que en general se tratan independientemente.

Líneas de investigación y desarrollo

El proyecto sigue la línea de investigación del campo de las bio-ciencias, en especial relacionada a la bioinformática, la bioingeniería y la biotecnología. En ese sentido, se ha comenzado a trabajar en la exploración de librerías en R para análisis cluster, en procesamiento distribuido en bioinformática utilizando el lenguaje r, en exploración e implementación de herramientas de desarrollo en un sistema de análisis estadístico de experimentos proteómicos. Todas estas líneas tienden a determinar los mejores y más efectivos métodos de análisis de la información procedente de datos biológicos, relacionados a agronomía, veterinaria, salud, ambiente.

Resultados y Objetivos

Objetivos:

Científicos

1. Extender los principios del Descubrimiento de Conocimiento en bases de datos y Minería de Datos en las Biociencias (Bioingeniería/Biotecnología/Bioinformática)
2. Estudiar y comparar métodos de clasificación/predicción basados en inteligencia artificial y en estadística sobre Bio-datos.
3. Formar recursos humanos nacionales en bioinformática.

Tecnológicos

4. Desarrollar herramientas y aplicaciones de bioinformáticas (librerías, software) para facilitar el procesamiento de datos biotecnológicos.
1. Editar un material de referencia sobre el uso de las aplicaciones desarrolladas.

Resultados:

Algunos integrantes disertaron en distintos Congresos Nacionales sobre temas relacionados a la temática propuesta, tal el caso de Laura Prato, quien disertó sobre "Aplicaciones bioinformáticas en experimentos con proteínas" en la X Semana de la Ciencia y la Tecnología en Villa María, en Junio 2012; y Cristobal Fresno dictó un curso sobre "Lenguaje R aplicado a Bioinformática" en el Workshops RSG-Argentina – Edición 2012, en Oro Verde, Entre Ríos en Septiembre 2012.

Se realizaron las siguientes Presentaciones en Congresos: en el 3er Congreso Argentino de Bioinformatica y Biologia Computacional, Oro Verde, Septiembre 2012 (Sesión de Posters), se presentaron los siguientes trabajos:

- GOboot: towards a robust SEA analysis. Cristobal Fresno, Andrea Llera, María Romina Girotti, María Pía Valacco, Juan A Lopez, Laura Zingaretti, Laura Prato, Osvaldo Podhajcer, Mónica G Balzarini, Federico Prada y Elmer Fernández.
- One vs One Artificial Neural Network strategy for gene expression multiclass classification. Remón L, Juarez L, Arab Cohen D, Fresno C, Prato L, Villoria L y Fernández E.
- SVM Tree with Optimal Multiclass Partition applied to Gene expression signature classification. Pellarolo M, Arab Cohen D, Fresno C, Prato L y Fernández E.

Y en las VII Jornadas de Investigación, Villa María, Noviembre 2012, se presentaron:

- Minería de datos en bio-ciencias: bioinformática, bioingeniería y biotecnología, Elmer A Fernández, Laura Prato, Cristóbal Fresno
- Exploración de distintas librerías y funciones de r para análisis cluster María Laura Zingaretti, Laura Prato, Elmer Fernández
- Procesamiento distribuido en bioinformática utilizando el lenguaje r, Mayco Fraccarolli, Emiliano Marzioni, Nicolás Ferreyra, Cristóbal Fresno, Laura Prato

Formación de Recursos Humanos

La estructura del equipo de trabajo tiene dos investigadores especialistas en la temática, un investigador encargado de la gestión, dos investigadores más relacionados al área técnico-informática y dos investigadores del área de desarrollo informático. También intervienen alumnos de grado en formación y para el desarrollo de sus respectivas tesinas.

Dos de los alumnos que intervienen han obtenido Beca para las Vocaciones Científicas, del CIN; y dos resultaron favorecidos con el Premio INTEL a la excelencia académica 2012 y 2013.

Referencias

• Gianola, D, Perez-Enciso M, Toro MA. 2003. On marker assisted prediction of genetic value: beyond the ridge. Genetics 163: 347–365.
• Gianola D, Fernando RL, Stella A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.
• Churchill, Doerge RW. 1994. Empirical Threshold Values for Quantitative Triat Mapping. Genetics 138: 963-971.
• Kang M, Balzarini M. Guerra J. 2004. Genotype-by-Environment interaction. In A. Saxton (ed.) SAS Genetic Book, SAS Institute, Cary NC (in press).
• 1-Buckingham S. (2003) Programmed for success. Nature. V425 pp 209-215
• Balzarini y cols (2004). Info-Gen: Software para análisis de datos genéticos. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Argentina. Copyright 362964 CESSI.
• U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996

• Fernández EA, Balzarini M. (2005) Biotecnología, Genómica y Minería de Datos. Enviado a Journal of Basic and Applied Genetics.

• Fernández EA, Balzarini M. (2005a) Improving Cluster Identification and Visualization in SOM like algorithms applied to Gene Expression Pattern Analysis. Enviado a BMC Bioinformatics.

• Fernández EA, (2005) Minería de Datos en Biotecnología. Aceptado en el Congreso Argentino de Bioingeniería 2005.

• Fernández EA. (2003) Minería de Datos y redes Neuronales para el análisis de la cinética de diálisis. Tesis de Doctorado, Universidad de Santiago de Compostela, España

• Fernández EA,Valtuille R, Willshaw P, Perazzo CA. (2001) Detection of Abnormality in the ECG Without Prior Knowledge by Using the Quantization Error of a Self-Organizing Map (SOM), tested on the European Ischemia Database, Medical & Biological Engineering & Computing. 39:330-337

• Cristóbal Fresno, Bioengineer; Andrea S Llera, Ph.D.; María R Girotti, Ph.D.; María P Valacco, Ph.D.; Juan A López; Osvaldo L Podhajcer, Ph.D.; Mónica G Balzarini, Ph.D.; Federico Prada, Ph.D.; Elmer A Fernández, The Multi-Reference Contrast Method: facilitating set enrichment analysis, Ph.D, Computers in Biology and Medicine. IN PRESS

• Fernandez EA, Souza Net EP, Abry P, Macchiavelli R, Balzarini M, Cuzin B, Baude C, Frutoso J, Gharib C Assessing erectile neurogenic dysfunction from heart rate variability through a Generalized Linear Mixed Model framework.. Computer Methods and Programs in Biomedicine doi:10.1016/j.cmpb.2009.11.001

• Fernández Elmer A, Girotti María R., López Juan A, Llera Andrea S., Podhajcer Osvaldo L, Cantet Rodolfo J. C. and Balzarini Mónica Improving 2D-DIGE protein expression analysis by two-stage linear mixed models: Assessing experimental effects in a melanoma cell study Bioinformatics

o Kohonen, T. (1997) Self-Organizing Maps (Springer, Berlin).

o Laciar E, Campos RJ, Brooks DH. Evaluación del daño miocárdico en enfermos chagásicos crónicos a partir del análisis del electrocardiograma de alta resolución… XV Congreso de Bioingeniería SABI 2005

• Podhajcer O. (2002). Estudios de genómica funcional mediante el uso de matrices de ADN (microarrays). Conferencia plenaria. Actas, XXXI Congreso Argentino de Genética, La Plata, 17-20 de setiembre de 2002.

• Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467-470.