Bio y AgroInformática en CIFASIS

Tapia, Elizabeth[1,2], Angelone, Laura[1,2], Bulacio, Pilar[1,2], Ornella, Leonardo[2], Coronel, José [1,2], Iglesias, Natalia [1,2], Murillo, Javier [2], Spetale, Flavio[2]

[1]Facultad de Cs. Exactas e Ingeniería, Av. Pellegrini 250, Rosario, Argentina
[2]CIFASIS-Conicet, Bv. 27 de Febrero 210 Bis, Rosario, Argentina

tapia@cifasis-conicet.gov.ar

Resumen

Las tecnologías de alto rendimiento en proyectos de ciencias de la vida generan cantidades exponenciales de datos cuya naturaleza y complejidad inspira el desarrollo de nuevos métodos computacionales para la extracción y gestión de información biológica relevante con el objetivo de lograr una comprensión más acabada de la vida tanto a nivel molecular como poblacional. Este contexto tecnológico, define un nuevo campo de investigación multidisciplinar conocido como **Bioinformática.**

En nuestro grupo estamos interesados en el desarrollo de algoritmos y herramientas bioinformáticas para el análisis, procesamiento y gestión de datos de espectroscopia, microarreglos, marcadores moleculares y de secuenciación de alto rendimiento en el marco de proyectos de investigación básica y biológica multidisciplinar.

Nuestro trabajo en Bioinformática inspira además la introducción de tecnologías de alto rendimiento y procesamiento de datos en **Agricultura de Precisión**, en el marco de un campo de investigación incipiente conocido como **Agroinformática**.

Palabras clave:
Bioinformática, Agroinformática

Contexto

Nuestros proyectos se llevan a cabo en el CIFASIS (Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas) en el marco de nuestro trabajo en la Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la UNR. Los mismos son financiados a partir de fondos provenientes de diferentes proyectos nacionales y provinciales, a saber:

PICT 2008 00253. *Clasificación de datos complejos mediante integración borrosa*. Agencia de Promoción Científica y Tecnológica. Secretaría de Ciencia y Técnica de la Nación. Período 2011-2014.

PICT 2006 02226. *Clasificación y Procesamiento Inteligente de la Información. Aplicaciones en Bioinformática, Agricultura de Precisión, Control y Comunicaciones.* Agencia de Promoción Científica y Tecnológica. Secretaría de Ciencia y Técnica de la Nación. Período 2007-2011.

Proyecto Nº 111910. *Desarrollo de un módulo electrónico versátil y económico para facilitar la transformación de protocolos SERIE-ETHERNET de uso industrial.* En colaboración con INGENEA SRL. Programa para la Promoción de la Vinculación Tecnológica entre el sistema productivo y el sistema de Ciencia y Tecnología en la provincia de Santa Fe 2010. Año de ejecución 2011.

Proyecto Nº 110309. *Desarrollo e implementación de un innovador dispositivo electrónico para conversión de protocolos en comunicaciones industriales.* En colaboración con INGENEA SRL. Programa para la Promoción de la Vinculación Tecnológica entre el sistema productivo y el sistema de Ciencia y Tecnología en la provincia de Santa Fe 2009. Año de ejecución 2010.

Red Sudamericana e Iberoamericana de Bioinformática. *E-Learning in Bioinformatics.* Programa Sul Americano de Apoio às Atividades de Cooperação em Ciência e Tecnología, PROSUL. Financiación CNPq nº 011/2008, Brasil. Período 2009-2011.

Estos proyectos involucran actividades de colaboración con investigadores del área biológica: el Dr. Esteban Serra del IBR (Instituto de Biología Molecular de Rosario, el Dr. Gerardo Cervigni del CEFOBI (Centro de Estudios Fotosintéticos y Bioquímicos) junto al Grupo de mejoramiento de frutales del EEA INTA San Pedro, y las Dras. Norma Paniego y Paula Fernandez del INTA Castelar.

Introducción

La generación de hipótesis, reglas, o modelos, a partir de datos es el nuevo paradigma de investigación científica por el que transitan actualmente muchas áreas de la ciencia, tanto

básica como aplicada. En este nuevo paradigma, la aplicación de técnicas de procesamiento inteligente es de fundamental importancia. Mediante estas técnicas, grandes cantidades de datos pueden ser analizadas de forma sistemática a los fines de recuperar información con la cual generar nuevas hipótesis, reglas, o modelos. Debe notarse, sin embargo, que las soluciones de procesamiento inteligente dependen en general del campo de aplicación y en muchos casos, del problema en sí. Debido a ello, su reutilización demanda el diseño de adaptaciones específicas y en casos extremos, el diseño de nuevas soluciones. En particular, se estima que una gran parte de las innovaciones y los desarrollos científicos de las próximas décadas, incluyendo la invención de sistemas computacionales más potentes, se inspirarán en soluciones encontradas al análisis de datos ómicos, datos cuya complejidad extrema demandará el desarrollo de una nueva generación de soluciones de procesamiento inteligente. En línea con estas argumentos, en nuestros proyectos se consideran problemas de recuperación de información en Bioinformática y Agricultura de Precisión derivados del uso de tecnologías de experimentación en Biología Molecular, y de la introducción de tecnologías de sensado de alta resolución para el monitoreo de procesos en Agricultura de Precisión.

La mayoría de nuestros trabajos se basan en problemáticas derivadas de experimentación en laboratorios genómicos (citados en Contexto) constituyendo así un trabajo mancomunado entre biólogos e informáticos. El diseño de herramientas informáticas en forma conjunta permite generar un lenguaje en común, clarificando y ordenando la información para la determinación de los algoritmos de solución. Además, el trabajo colaborativo permite la identificación de intereses comunes y la planificación de acciones conjuntas con las que abordar, de manera multidisciplinaria, proyectos de investigación, y contribuir a la consolidación de un área de vacancia como lo es la Bioinformática en nuestro país.

La Agricultura de Precisión trata con los métodos, las tecnologías, y las estrategias de gestión aplicables al estudio y tratamiento de parámetros en los procesos de cultivo que, debido a variabilidad espacial o temporal, influyen en el rendimiento y sustentabilidad de cultivos así como en la protección del medio ambiente. Una tecnología importante en Agricultura de Precisión es la de los monitores de siembra. Una de las funciones básicas de los monitores de siembra es controlar el estado de cada línea de siembra y la densidad de siembra. Esta información es relevante principalmente para cultivos de maíz (Zea mays L.) y girasol (Helianthus agnus). Mediante un procesamiento inteligente, la información de siembra de alta precisión podría usarse para certificar la calidad de sembradoras comerciales. En la búsqueda de una solución tecnológica para la medición de densidad de siembra de alta resolución, el grupo ha propuesto un sistema electrónico de sensado de alta resolución (paralelo y masivo) de la densidad siembra mediante la introducción de un procesamiento distribuído en líneas de siembra. A partir de la aplicación de esta tecnología y con el objetivo de analizar el patrón de variabilidad espacial de la densidad de siembra en cultivos de maíz y girasol, se propuso el desarrollo de un modelo de variabilidad espacial de la densidad de siembra con el objetivo principal de lograr estimadores de rendimiento más precisos. Lo cual pueden ser usado para: i) analizar la eficiencia de las maquinas sembradoras, ii) modelar de forma robusta la variabilidad espacial de la densidad de siembra, y iii) estimar, de forma más precisa, el rendimiento esperado, factor fundamental en la toma de decisiones de naturaleza económica.

Líneas de investigación y desarrollo

Bioinformática

- Diseño de clasificadores multiclase basados en códigos correctores de error para el análisis de datos biológicos (datos con ruido y/o alta dimensionalidad) [1-8-9-15]
- Diseño de algoritmos de selección automática de variables en datos biológicos mediante medidas de conjunto (integrales borrosas) [10]
- Análisis, adaptación y desarrollo de herramientas para el análisis de datos biológicos de diversa naturaleza (secuencia, expresión, marcadores, espectroscopia, etc.) con énfasis en el uso de técnicas del Aprendizaje de Máquina. [7-13-14]
- Diseño de bases de datos en proyectos de genómica funcional. [5-6-11-12]

Agroinformática

- Desarrollo de tecnologías (hardware y software) ISOBUS para el sensado masivo de variables de interés agronómico. [2-3-4]

Resultados y Objetivos

En el área de Bioinformática se diseñaron algoritmos de clustering basados en métricas de grafos, de clasificación de datos no estacionarios,

de fusión de clasificadores heterogéneos con integrales fuzzy, de clasificacion multiclase con classificadores binarios usando códigos. Para su prueba y verificación se utilizaron datos simulados y datos biológicos reales caracterizados por una alta dimensionalidad, desbalance y ruido en las muestras.

Se desarrolló una herramienta Java para la determinación de sitios de inicio de transcripción en genomas de Tripanosomas a partir de clasificadores AdaBoost.

Se diseño e implementó una base de datos para almacenar, gestionar y consultar información genómica de ESTs de girasol dentro del marco de un proyecto llevado a cabo en INTA Castelar.

En el área de Agricultura de Precisión se diseñó y puso a prueba un sistema de sensores optoelectrónicos con microprocesadores integrados para la medición de alta resolución de la distribucion espacial de siembra. Las pruebas de laboratorio de sensores individuales permitieron establecer las velocidades de siembra y posiciones de montaje más convenientes. El prototipo de este sistema fue distinguido con Medalla de Oro en el rubro Siembra en la edición 2009 del Premio Ternium Siderar Expoagro a la Innovación en Maquinaria Agrícola.

Los resultados forman parte de los trabajos publicados que se detallan en Referencias.

Premios

Coronel J., Tapia E. (2009) *Sensor de Densidad de Siembra.* Medalla de Oro en el rubro Siembra en la edición 2009, premio Ternium Siderar Expoagro a la Innovación en Maquinaria Agrícola; con el apoyo de la Sociedad Alemana de Agricultura DLG (Deutsche Landwirtschafts-Gesellschaft), institución organizadora de la exposición agrícola AGRITECHNICA1 desde el año 1985 y del Premio Agritechnica Neuheiten desde hace más de 40 años.

Formación de recursos humanos

Tesis de Posgrado

A continuación se detallan las tesis doctorales que se desarrollan en el grupo.

MSc.Leonardo Ornella. Tesis de Doctorado: *Códigos Correctores de Error en Problemas de Clasificación Multiclase de Datos de Marcadores Moleculares*. (finalizada)

Ing. José Coronel. Tesis de Doctorado: *Diseño de un sistema de alta resolución para la estimación de la distancia entre semillas.* (en curso)

Ing. Natalia Iglesias. Tesis de Doctorado: *Desarrollo de tecnologías de agricultura de precisión sobre ISOBUS.* (en curso)

Lic. Javier Murillo. Tesis de Doctorado: *Desarrollo de metaclasificadores FI para clasificación de datos complejos.* (en curso)

Ing. Flavio Spetale. Tesis de Doctorado: *Metaclasificadores FI para datos espectrales asociados al reciclaje de polímeros.* (en curso)

Cursos de Posgrado

Desde el año 2003 se promueve y colabora en la difusion de la Bioiformatica. Se colaboró en el dictado de cursos de Doctorados abierto a la comunidad científica.

Recientemente se dictaron los siguientes cursos:

Bioinformática. Aspectos estadísticos del análisis de datos de microarreglos. A cargo de la Dra. Diana M. Kelmansky (UBA). Duración 40 horas. Junio 2010, Laboratorio de Computación de FCEIA-UNR, Avda. Pellegrini 250.

Bioinformática. Estadística aplicada a estudios experimentales. A cargo del MSc. Julio Di Rienzo (UNC). Duración 40 horas. Agosto 2010.

Laboratorio de Computación de FCEIA-UNR, Avda. Pellegrini 250.

Referencias **(últimos 2 años)**

[1] Tapia E., Ornella L., Bulacio P., Angelone, L. (2011). Multiclass classification of microarray data samples with a reduced number of genes. BMC Bioinformatics, 12:59. doi: 10.1186/1476-2105-12-59. Publisher BioMed Central.

[2] Coronel J., Ornella L., Bulacio P., Nardón G., Tapia E. (2011) Testing of an Opto-Electronic Sensor for the high throughput measurement of seed spatial distributions. L.A.A.R.: Latin American Applied Research - An International Journal, Vol. 42:2.

[3] Iglesias, N.; Coronel, J.; Bulacio, P.; Tapia E. (2010) Baja Distorsión en la reconstrucción de Distribución Espacial de Siembra de Alta Resolución. Congreso Brasileño de Agricultura de Precisión - ConBAP. Septiembre 2010. Ribeirao Preto-SP, Brasil. In Press.

[4] Iglesias, N.; Coronel, J.; Tapia E. (2010) Tasa de Muestreo Óptima en Medición de Alta Resolución de Distribución Espacial de Semillas. IX Congreso Latinoamericano y del Caribe de Ingeniería Agrícola. XXXIX Congresso Brasileiro de Engenharia Agrícola - CANBEA. Julio 2010. Vitória-ES, Brasil. In press.

[5] A Pons, C Reynares, L Angelone, P Fernández, P Bulacio, N Paniego, E Tapia (2010), Evolución de la Interfaz de Consulta de la SfGD. Un Puente de Entendimiento Informático-Biológico, 39 JAIIO-CAI. In press.

[6] Fernández, P, Blesa, D, Príncipi, D, Fusari, C, Soria, M, Reynares, C, Angelone, L, Delfino, S, Conesa, A, Dopazo, J, Tapia, E, Heinz, R and Paniego, N (2010) Sunflower Functional Genome Database, a curated unigene database to support functional diversity studies en sunflower, ISCB Latin America, Uruguay.

[7] Ornella L., Bulacio P., Tapia E. (2010) Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. Computers and Electronics in Agriculture, Vol 74(2):250-257. Publisher: Elsevier.

[8] Ornella L., Tapia E. (2010). Application of error correcting codes for heterotic group assignation. Maize Genetics Cooperation Newsletter Vol. 84:1-2.

[9] Tapia E., Bulacio P., Angelone L. (2009) Recursive ECOC Classification. Pattern Recognition Letters. Vol. 31(3):210-215. Publisher: Elsevier.

[10] Bulacio P., Guillaume S. Tapia, E., Magdalena L. (2009). A selection approach for scalable fuzzy integral combination. Information Fusion. Vol. 11(2):208-213. Publisher: Elsevier.

[11] Angelone L., Ornella L., Bulacio, P., Tapia E. (2009) GibbsSM: Predicción Automática de Motivos mediante Muestreo Gibbs, Jornadas de Ciencia y Tecnología-Divulgación de la Producción Científica y Tecnológica de la UNR, Diciembre 2009. In press.

[12] Fernández P., Angelone L., Bulacio P., Reynares C., Tapia E., Paniego N. (2009). SfGD: Base de Datos Genómicos de Girasol, Congreso de Agro informática 2009, CAI 2009, 24 al 28 de agosto 2009, ISSN 1852-4850.

[13] Tapia, E., Esteban, L., Bulacio P., Angelone L., Serra E. (2009). MET: Methionine Exploration of Trypanosomes Software, RPIC 2009, XIII Reunión de Trabajos en Procesamiento de la Información y Control, Septiembre 2009, Rosario, Argentina. ISBN 950-665-340-2.

[14] Ornella L., Tapia E. (2009) Applications of Machine Learning in Ecological Modeling. Mathematical Modeling of Biophysical Phenomena Meeting, Angra Dos Reis, Brazil, March 2009.

[15] Ornella L., Tapia E. (2009) Applications of error correcting codes in assigning new maize (Zea mays L.) inbreds to known heterotic groups using molecular marker information. II Congreso Internacional-REDBIO-Argentina. Rosario, Arg., Abril 2009.