

Bioinformática: herramienta en el análisis de sistemas biológicos

Giovanni Almanzar*

Diana Marcela González Ciro**

Fecha de recepción: 13 de septiembre de 2013

Fecha de aprobación: 20 de octubre de 2013

Pp. 25 a 38

RESUMEN

Bioinformática es la ciencia que permite adquirir, almacenar, organizar, archivar, analizar o visualizar datos complejos relacionados en su mayoría con sistemas biológicos. La bioinformática requiere de herramientas de *Software*, que permitan el manejo de datos, la creación de modelos, la evaluación de hipótesis y la interpretación de los resultados con sentido biológico. Este artículo describe algunas de las herramientas y *Software* empleados en bioinformática.

PALABRAS CLAVE

Bioinformática, modelo computacional, *Software*, sistemas biológicos.

* PhD. en Bioquímica, Universidad de Salzburgo, Austria.

** Ingeniera de Sistemas, Universidad de Manizales.

Bioinformatics: a tool for the analysis of biological systems

ABSTRACT

Bioinformatics is a science which allows investigators to acquire, organize, access, store, analyze or visualize complex data mostly related with biological systems. Bioinformatics requires software tools that allow data management, hypothesis evaluation and interpretation of biological outcomes. Then, this article describes some tools and software used in Bioinformatics.

KEY WORDS

Bioinformatics, Computational model, Software, Biological Systems.

La Bioinformatique: un outil d'analyse des systèmes biologiques

RÉSUMÉ

La bio-informatique est la science qui permet d'acquérir, d'organiser, d'archiver, d'analyser des données complexes traitant essentiellement de systèmes biologiques. La bioinformatique exige des logiciels informatiques permettant la gestion de données, la création de modèles, l'évaluation d'hypothèses et l'interprétation de résultats. Cet article décrit certains des outils et des logiciels utilisés en bio-informatique.

Mots-clés

Bioinformatique, Modèle Informatique, Logiciels, Systèmes Biologiques.

Bioinformática: uma ferramenta para a análise de sistemas biológicos

RESUMO

A bioinformática é a ciência que permite adquirir, armazenar, organizar, arquivar, analisar ou visualizar dados complexos relacionados maioritariamente aos sistemas biológicos. A bioinformática precisa de ferramentas de software que permitam o tratamento de dados, a criação de modelos, a avaliação da(s) hipóteses e a interpretação dos resultados com sentido biológico. Este artigo descreve algumas das ferramentas e software empregados em bioinformática.

PALAVRAS-CHAVE

Bioinformática, Modelo Computacional, Software, Sistemas Biológicos.

1. Introducción

El avance de la ciencia junto con el avance tecnológico, ha generado una inmensa cantidad de información que año tras año y desde finales de los años 80 crece exponencialmente. Esta información o estos datos fundados por investigaciones científicas en ramas como la biología, química, física y matemáticas, entre otras, deben ser organizados, procesados, analizados e interpretados. Herramientas tales como la informática y la estadística tienen un papel primordial en dicho procedimiento, dando paso a lo que hoy conocemos como bioinformática. Es así como, más de 90.000 publicaciones registradas en PubMed están relacionadas con este tema.

Pero, ¿Qué es la bioinformática? El *National Institutes of Health (NIH)*, la definió como la investigación, desarrollo o aplicación de herramientas computacionales para expandir el uso de datos biológicos, médicos, comportamiento o salud, incluyendo aquellos empleados para adquirir, almacenar, organizar, archivar, analizar o visualizar dichos datos. Y definió biología computacional, como el desarrollo y aplicación del análisis de datos y métodos teóricos, modelos matemáticos y técnicas de simulación computacional para el estudio de sistemas biológicos, de comportamiento y social (Huerta, 2000) Términos frecuentemente empleados como sinónimos.

Para Luscombe y colaboradores, bioinformática es la conceptualización biológica en términos de macromoléculas (en el sentido fisicoquímico) y la aplicación de técnicas informáticas (derivadas de disciplinas como las matemáticas, ciencia computacional y estadística) para entender y organizar la información asociada con estas moléculas a gran escala (Luscombe et al., 2001). De manera un poco más sencilla, se definiría como la aplicación de tecnología computacional al manejo y análisis de datos biológicos en el cual los computadores son empleados para reunir, almacenar, analizar y fusionar datos biológicos.

Como objetivos, la bioinformática plantea la organización de datos, de tal forma que permita a los investigadores acceder a la información existente y registrar nuevas entradas; desarrollar herramientas y recursos que ayuden en el análisis de datos. El desarrollo de dichos recursos requiere de conocimientos en teoría computacional, así como la interpretación biológica de los datos analizados; usar herramientas para analizar e interpretar los resultados con sentido biológico (Luscombe et al., 2001). En esencia, la bioinformática hace un análisis global del total de datos disponibles, a fin de descubrir principios comunes dentro de diferentes áreas y que puedan resaltar nuevas características.

El avance tecnológico aplicado a diferentes áreas de la ciencia, ha potenciado la generación de una gran cantidad de información que trata de explicar cómo ocurren los procesos biológicos en la naturaleza. El sufijo *oma* que significa conjunto, ha sido asignado para definir diferentes poblaciones y subpoblaciones en la célula y *omica* refiere al tópico de investigación (Greenbaum et al., 2001)

2. Herramientas empleadas en bioinformática

Los sistemas biológicos fueron creados para lograr el entendimiento de procesos en organismos vivos y aplicar este conocimiento a diferentes campos como la medicina y la biotecnología (Ideker et al., 2001; Kitano, 2002b; Kitano, 2002a). Avances significativos se han obtenido en el análisis de vías de expresión, señalización celular y desarrollo de medicamentos por ejemplo, contra el cáncer.

Las herramientas computacionales son de extrema utilidad, no solo en sistemas biológicos, sino adicionalmente en áreas industriales como la aviación, el entretenimiento y la electrónica (Evans, 2006). Las herramientas de *Software*, buscan facilitar el trabajo de los investigadores en diferentes áreas, ayudan en la búsqueda, prevención y solución de enfermedades y en la solución a diversas preguntas que se presentan día a día (Ghosh et al. 2011).

A continuación se describen las diferentes herramientas empleadas en bioinformática.

2.1 Manejo de datos (*Data Management*)

La adquisición y manejo de datos es de vital importancia en la generación y evaluación de una hipótesis. Para el manejo de grandes volúmenes, se han creado numerosos *Data Management* (Field et al., 2005; Keator, 2009). Sin embargo, estos sistemas de manejo necesitan de formatos estándar para el intercambio e identificación de datos únicos en el mapeo de la información (*data mapping*) y de interfaces comunes que permitan la integración de estos en el proceso computacional (Van Deun et al., 2009).

2.1.1 Manejo de datos (*Data Management Standards*)

Enfocados en tres aspectos básicos:

- **Información mínima:** es una lista de control de la información de soporte del conjunto de datos de diferentes experimentos. Algunos ejemplos son: *Minimum Information About a Microarray Experiment* (MIAME) (Brazma et al., 2001); *Minimun Information About a Proteomic Experiment* (MIAPE) (Martens et al., 2008) y *Minimun Information for Biological and Biomedical Investigation* (MIBBI) (Taylor, 2008).
- **Formato de archivos:** define la información que debe ser almacenada. Los formatos están generalmente basados en *Extensible Markup Language* (XML), que facilita procesos automatizados por computadores.
- **Ontologías:** definen la relación y la jerarquía entre términos y permite una única semántica en la anotación de los datos. Como ejemplos de ontología encontramos *Gene Ontology* (GO) y *Systems Biology Ontology* (SBO).

2.1.2 Manejo de datos y herramientas para el análisis de datos (*Data Management and Data Analysis Tools*)

Tablas, páginas *Web* o sistemas de manejo de datos de laboratorio son las formas comunes de clasificación. Programas tabulares son los más amplios y comúnmente empleados para el almacenamiento de datos y comunicación entre la comunidad científica. Sin embargo, la integración de estos programas con herramientas de análisis y procesos computacionales, requiere de interfaces que no son siempre soportadas por diferentes *Softwares*. El uso de sistemas de manejo de procesos (*Workflow-Management Systems*, WMSs) brinda el primer paso para construir una línea computacional, habilitando el intercambio e integración de datos y la comunicación entre herramientas. Sin embargo, estos sistemas solo soportan conjuntos específicos de herramientas y estándares, forzando el uso de nuevos WMSs para el entendimiento de los sistemas biológicos de interés (Ghosh, et al.).

2.2 Red de inferencia de datos

La red de inferencia de datos (*Data-driven Network-based Modelling*), emplea algoritmos para inferir relaciones dentro de entidades moleculares como genes, proteínas, etc., que han sido obtenidos bajo diferentes condiciones experimentales. El modelo resultante, se conoce como red de inferencia, red de co-expresión o red de asociación. El enfoque de las redes de inferencia de datos, está basado en evaluar la probabilidad de una hipótesis con la observación o la evidencia experimental. Técnicas alternativas como regresión, correlación de métodos y el empleo de información común, han sido aplicadas en la evaluación de la hipótesis (Margolin, et al., 2006; Faith, et al., 2007; Fadda, et al., 2009). El principal reto de las técnicas de redes de inferencia esta en solucionar la sub-determinación, donde el número de posibles interacciones inferidas excede el número de medidas independientes.

2.3 Depuración de datos (*Deep Curation*)

La depuración de datos crea un mapa de interacción molecular detallada para una integración a gran escala del conocimiento, por ejemplo información de publicaciones, bases de datos y el flujo de datos (Caron et al.). La depuración de datos, crea un modelo manual o semimanual haciéndolo de fácil manejo para investigadores, en el cual se pueden adicionar hipótesis propias. Los usuarios pueden agregar interacciones desconocidas como hipótesis dentro de la vía de depuración. Estas solo podrán ser beneficiosas si se incluye un soporte racional de la hipótesis.

La depuración de datos provee detalles sobre el mecanismo de cada interacción, permitiendo la observación de los reportes experimentales y la literatura críticamente. En investigación básica y desarrollo de medicamentos, es esencial cuando la prioridad es entender los detalles de los mecanismos moleculares más allá de identificar nuevas moléculas o interacciones.

Algunos *Software* empleados para depuración de datos son: *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa and Goto, 2000);

Reactome (Joshi-Tope, et al., 2005); *Panther Pathway Database* (Mi et al., 2005), *Pathway Commons* (Cerami et al.) y *BioCyc* (Karp et al., 2005). Otros modelos son: *Systems Biology Markup Language* (SBML) (Hucka et al., 2003) y *Biological Pathways exchange* (BioPAX) (Demir, 2010).

- **Estándares y herramientas para simulación**

Las simulaciones tienen un papel importante en la verificación computacional de modelos biológicos y en la predicción computacional de comportamientos. Después de creado el modelo en base a las hipótesis, simulaciones dinámicas evalúan el modelo con el fin de saber si este se comporta como el sistema biológico real. Algunos ejemplos de simuladores son: SBML, *Systems Biological Graphical Notation* (SBGN) (Le Novère, 2009) y *Minimum Information Required in the Annotation of Models* (MIRIAM) (Le Novère, 2005)

- **Análisis de modelos**

Muchas de las técnicas de análisis de modelos, se enfocan en la dinámica del sistema que está representado como un conjunto de ecuaciones diferenciales ordinarias. Análisis alternativos han sido desarrollados con base a redes de análisis estadístico (Hofestadt y Thelen, 2011) .

2.4 Plataforma integrada de *Software*

Han sido creadas para encaminar esfuerzos de productividad, mejoramiento de calidad e innovación en la industria (Evans, 2006). El concepto detrás de la creación de plataformas integradas de software, consiste en habilitar usuarios para acceder a datos y conocimiento desde cualquier punto en el diagrama de flujo que permita la adaptación de este a las necesidades del usuario, incrementando los niveles de interoperatividad. Estos factores, reducen costos asociados con el uso independiente e incompatible de *Software*. A consecuencia, los usuarios emplean más tiempo y esfuerzo en generar ciencia que en la operación de *Software*.

3. Conclusión

El desarrollo e integración rápida y efectiva de la información obtenida de diversos análisis de sistemas biológicos con la informática, brindan una amplia y fascinante área del conocimiento donde nunca más el estudio particular de una vía de expresión, función de una proteína o interacción de esta con otras moléculas en un organismo, será tomada aisladamente. Al contrario, la bioinformática y sus diversas herramientas computacionales, permiten el estudio de la molécula en un entorno general. La bioinformática es una valiosa herramienta en el diseño, producción y evaluación de procesos industriales como el empleado en el desarrollo de medicamentos para el tratamiento contra diversas enfermedades.

4. Referencias bibliográficas

- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data*. Nat Genet 29, 365-71.
- Caron, E., Ghosh, S., Matsuoka, Y., Ashton-Beaucage, D., Therrien, M., Lemieux, S., Perreault, C., Roux, P.P. and Kitano, H. (2010). *A comprehensive map of the mTOR signaling network*. Mol Syst Biol 6, 453.
- Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D. and Sander, C. (2011). *Pathway Commons, a web resource for biological pathway data*. Nucleic Acids Res 39, D685-90.
- Demir, E. (2010) *The BioPAX community standard for pathway data sharing*. Nature Biotech. 28, 735-741.
- Evans, D., Hagi, A., Schmalensee, R. (2006) *Invisible Engines: How software Platforms Drive Innovation and Transform Industries*.
- Fadda, A., Fierro, A.C., Lemmens, K., Monsieurs, P., Engelen, K. and Marchal, K. (2009) *Inferring the transcriptional network of Bacillus subtilis*. Mol Biosyst 5, 1840-52.
- Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, T.S. (2007) *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles*. PLoS Biol 5, e8.
- Field, D., Tiwari, B. and Snape, J. (2005) *Bioinformatics and data management support for environmental genomics*. PLoS Biol 3, e297.

- Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.Y. and Kitano, H. (2011). *Software for systems biology: from tools to integrated platforms*. Nat Rev Genet 12, 821-32.
- Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J. and Gerstein, M. (2001) *Interrelating different types of genomic data, from proteome to secretome: 'oming in on function*. Genome Res 11, 1463-8.
- Hofestadt, R. y Thelen, S. (2011). *Quantitative modeling of biochemical networks*. Stud Health Technol Inform 162, 3-16.
- Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. and Wang, J. (2003) *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*. Bioinformatics 19, 524-31.
- Huerta, M., Downing, G., Haseltine, F., Seto, B., and Liu, Y. (2000). *NIH working definition of bioinformatics and computational biology*. Bioinformatics Definition Committee, NIH.
- Ideker, T., Galitski, T. and Hood, L. (2001). *A new approach to decoding life: systems biology*. Annu Rev Genomics Hum Genet 2, 343-72.
- Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005). *Reactome: a knowledgebase of biological pathways*. Nucleic Acids Res 33, D428-32.
- Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30.
- Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. and Lopez-Bigas, N. (2005). *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes*. Nucleic Acids Res 33, 6083-9.

- Keator, D.B. (2009). *Management of information in distributed biomedical collaboratories*. Methods Mol Biol 569, 1-23.
- Kitano, H. (2002a). *Computational systems biology*. Nature 420, 206-10.
- Kitano, H. (2002b). *Systems biology: a brief overview*. Science 295, 1662-4.
- Le Novere, N.(2005). *Minimum information requested in the annotation of biochemical models (MIRIAM)*. Nature Biotech. 23, 1509-1515.
- Le Novere, N. (2009). *The Systems Biology Graphical Notation*. Nature Biotech. 28, 935-942.
- Luscombe, N.M., Greenbaum, D. and Gerstein, M. (2001). *What is bioinformatics? A proposed definition and overview of the field*. Methods Inf Med 40, 346-58.
- Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006). *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. BMC Bioinformatics 7 Suppl 1, S7.
- Martens, L., Palazzi, L.M. and Hermjakob, H. (2008). *Data standards and controlled vocabularies for proteomics*. Methods Mol Biol 484, 279-86.
- Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H. and Thomas, P.D. (2005). *The PANTHER database of protein families, subfamilies, functions and pathways*. Nucleic Acids Res 33, D284-8.
- Taylor, C.F.e.a. (2008). *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project*. Nature Biotech. 26, 889-896.
- Van Deun, K., Smilde, A.K., van der Werf, M.J., Kiers, H.A. and Van Mechelen, I. (2009). *A structured overview of simultaneous component based data integration*. BMC Bioinformatics 10, 246.