Licenciatura en
CIENCIAS
GENÓMICAS

Formato para proponer cursos
Semestre __2023-2_____

El curso ya ha sido impartido: Sí__**X**__ No____

1. Indicar modalidad: (Optativo, Seminario, curso regular (con profesor invitado) Curso regular
2. Título: Se sugiere que sea conciso y refleje el contenido general **Bioinformática y bioestadística II**: análisis de algoritmos en bioinformática usados en el análisis de motivos de DNA, análisis de secuencias de HT, de métodos de clustering, análisis de macromoléculas y procesamiento de lenguaje natural.
3. Tutor responsable:

Nombre completo	**Julio Collado y Heladia Salgado**

6. Descripción del curso

El objetivo del curso es comprender a profundidad los principales algoritmos usados en bioinformática para: el análisis de motivos, análisis de secuencias generadas por métodos high throughput (HT), métodos de clustering, métodos de procesamiento de lenguaje natural en genómicas y análisis de macromoléculas.

Introducción al Uso del Cluster
Profesor: Alfredo Hernández (AH)
Objetivo: Repasar las funcionalidades del sistema operativo unix.

- Comandos para el manejo de procesos
- Comandos para el manejo del cluster

Módulo: Análisis de datos de secuenciación masiva
Profesor: Leonardo Collado Torres (LCT)
Objetivo: Durante el curso los alumnos recibirán una introducción a las tecnologías de secuenciación utilizadas en la actualidad, así como la gran variedad de experimentos que con ellos pueden realizarse. Entenderán las ventajas y limitantes que presenta esta tecnología. En la parte práctica, utilizarán salidas reales de secuenciadores actuales. Evaluarán la calidad de los datos y los filtros que conviene usar para optimizar los datos para diferentes aplicaciones. Aprenderán a usar R/Bioconductor y diversos programas en Linux para procesar los datos y hacerles preguntas de relevancia biológica. Específicamente realizarán análisis de calidad, ensamble de novo, mapeo, anotación y expresión diferencial.

Módulo: Algoritmos de búsqueda y descubrimiento de motivos de DNA
Profesor: Jacques van Helden (JVH)
Objetivo: Se explican los algoritmos y la estadística detras de las herramientas de análisis de secuencias de DNA. Especificamente se revisarán algoritmos para la búsqueda y descubrimiento de sitios de unión de reguladores

transcripcionales. Durante el curso se utilizarán bases de datos de regulación y la herramienta Regulatory Sequence Analysis Tools (http://embnet.ccg.unam.mx/rsa-tools/).

Módulo: Algoritmos en bioinformática estructural
Profesor Invitado: Dr. José Arcadio Farías Rico (JFR)
Objetivo: Conocer los principales algoritmos para el estudio y predicción de la estructura de proteínas y ADN.

Módulo: Algoritmos de agrupamiento de datos (Clustering)
Profesor: Arturo Medrano Soto (AMS)
Objetivo: Conocer los métodos de agrupamiento de datos así como sus ventajas y desventajas.

Módulo: Algoritmos de aprendizaje supervisado
Profesor: Carlos-Francisco Méndez-Cruz (CMC)
Objetivo: el alumno reconocerá las principales técnicas y métodos de procesamiento de lenguaje natural para el análisis automático de literatura científica en biomedicina. Además, implementará un pipeline para clasificación automática de artículos científicos.

Módulo: Introducción a la Ciencia de Datos
Profesor: Carlos-Francisco Méndez-Cruz (CMC)
Objetivo: Al finalizar el curso, el alumno será capaz de aplicar diversas técnicas de análisis de datos, visualización y programación para resolver un problema relacionado con las ciencias genómicas a partir de un conjunto de datos.

Planificación y Organización
Julio Collado, Heladia Salgado

- Revisión de retroalimentación del curso por parte de la dirección de la LCG
- Revisión de temas de vanguardia en bioinformática para actualizar el contenido del curso.
- Revisión y discusión con investigadores colaboradores sobre el Contenido del Curso
- Búsqueda y selección de profesores invitados
- Selección de ayudantes.
- Presentación del Plan del Curso a profesores invitados y ayudantes.
- Solicitud de infraestructura computacional para el curso y sus módulos.
- Logística para traer profesores invitados extranjeros expertos en el área, cuando son clases presenciales.
- Seguimiento del curso junto con los ayudantes
- Calificar Acta del curso
- Solicitud de constancias de participación
- Lecciones aprendidas del curso

7. Características para la impartición del curso :	
Lugar donde se realizará	Licenciatura en Ciencias Genómicas
Duración en horas por sesión y número de sesiones	116 horas 4 horas diarias de lunes a viernes por 4 semanas (80 hrs.) 12 sesiones de 3 horas del módulo de Ciencia de Datos (36 hrs.)
Disponibilidad de impartirlo por videoconferencia	Sí_X_ No____
8. Método de evaluación:	
Por favor incluya en este apartado el % de la contribución relativa de:	
Participación en clase	
Presentación en clase	

Proyecto de investigación	
Trabajos	
Otros	Cada módulo tiene un porcentaje igual para cubrir entre todos el 100%. La calificación final del alumno, se obtiene sumando los porcentajes obtenidos en cada módulo. Cada Profesor del módulo indicará la forma de evaluación, pero en general se toma en cuenta la participación, las tareas y proyectos.

10. Bibliografía

Referencias:

1. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. 2015;16:153-72. doi: 10.1146/annurev-genom-090314-050032. Epub 2015 Apr 22. Review. PubMed PMID: 25939056.
2. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011 Nov 8;29(11):987-91. doi: 10.1038/nbt.2023. PubMed PMID: 22068540.
3. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525-7. doi: 10.1038/nbt.3519. Epub 2016 Apr 4. PubMed PMID: 27043002.
4. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015 Feb;12(2):115-21. doi:10.1038/nmeth.3252. Review. PubMed PMID: 25633503; PubMed Central PMCID: PMC4509590.
5. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494-512. doi: 10.1038/nprot.2013.084. Epub 2013 Jul 11. PubMed PMID: 23845962; PubMed Central PMCID: PMC3875132.
6. van Helden J. Regulatory equence analysis tools. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6.
7. van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 2000 Apr 15;28(8):1808-18.
8. Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genesby computational analysis of oligonucleotide frequencies. J Mol Biol. 1998 Sep 4;281(5):827-42.
9. Stormo GD. Consensus patterns in DNA. Methods Enzymol. 1990;183:211-21.
10. Hertz GZ, Hartzell GW 3rd, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci. 1990Apr;6(2):81-92.
11. Schug J, Overton GC. Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. Proc Int Conf Intell Syst Mol Biol. 1997;5:268-71.
12. Zhou Q, Liu JS. Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics. 2004 Apr 12;20(6):909-16. Epub 2004 29.
13. Bailey TL, Baker ME, Elkan CP. An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. J Steroid Biochem Mol Biol. 1997 May;62(1):29-44.
14. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 2002 Nov;11(11):2606-21.
15. Contreras-Moreira B, Branger PA, Collado-Vides J. TFmodeller: comparative modelling of protein-DNA complexes. Bioinformatics. 2007 Jul 1;23(13):1694-6. Epub 2007 Apr 25
16. Ananiadou, S., & McNaught, J. (2006). Text mining for biology and biomedicine. Boston: Artech House.
17. Gama-Castro, S., Rinaldi, F., López-Fuentes, A., Balderas-Martínez, Y. I., Clematide, S., Ellendorff, T. R., ... & Collado-Vides, J. (2014). Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12. Database, 2014.
18. Huang, C. C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics.
19. Jurafsky, D. & Martin, J. H. (2007). Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, N.J.: Pearson Prentice Hall.
20. Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). Text mining: predictive methods for analyzing unstructured information. Springer Science & Business Media.

21. Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. Cluster Analysis. 5th Edition. 2011, John Wiley & Sons, Lt. ISBN: 978-0-470-74991-3
22. Stephane Tuffery. Data Mining and Statistics for Decision Making. First Edition. 2011, John Wiley & Sons, Ltd. ISBN: 978-0-470-68829-8.
23. Jain A. K. Murty M. N. Flynn P. J. Data Clustering: a Review. 1999, ACM Computing Surveys 31(3):264-323
24. Cady, F. (2017). The data science handbook. John Wiley & Sons.
25. Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.
26. Igual, L., & Seguí, S. (2017). Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications. Springer, Cham.
27. Pathak, M. A. (2014). Beginning data science with R. Springer.
28. Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media.
29. Mailund, T. (2017). Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist. Apress.
30. Ismay, C & Kim, A. Y. (2022). Statistical Inference via Data Science. A ModernDive into R and the Tidyverse. CRC Press. Versión digital en https://moderndive.com/index.html.
31. Mardia, K. V., Kent, J. T. & Bibby, J. M. (2008). Multivariate analysis. Academic Press.
32. Basel Abu-Jamous, Rui Fa, Asoke K. Nandi. Integrative Cluster Analysis in Bioinformatics. First Edition. 2015, John Wiley & Sons, Lt. ISBN: 978-1-118-90653-8