

NOMBRE: **BIOINFORMÁTICA (B1)**

PROFESOR RESPONSABLE: **Enrique Viguera Mínguez (Universidad de Málaga)**

PROFESORES PARTICIPANTES

Claros Díaz, M. Gonzalo (Facultad de Ciencias. Univ. de Málaga)

Trelles Salazar, Oswaldo (ETSI de Informática. Univ. De Málaga)

Thode Mayoral, Guillermo (Facultad de Ciencias. Univ. de Málaga)

Pérez Pulido, Antonio J. (Facultad de Ciencias. Universidad Pablo Olavide)

Ugo Bastolla (Centro de Biología Molecular “Severo Ochoa” (CSIC-UAM))

DESCRIPTORES

Introducción a la Bioinformática. Bases de datos de secuencias de ADN y proteínas. Búsquedas de bases de datos mediante palabras clave. Búsqueda de homologías mediante alineamiento local. Análisis básico de secuencias de ADN y proteínas. Alineamientos global y múltiple. Filogenia molecular. Proyectos Genoma. Genómica funcional: predicción computacional de función. Genómica comparada: herramientas para la comparación de genomas completos. Procesamiento de datos de microarrays y métodos de clasificación.

OBJETIVOS DEL APRENDIZAJE

El progresivo desarrollo de métodos automatizados de preparación de muestras de DNA y secuenciación ha permitido afrontar, a lo largo de la última década, diversos proyectos de secuenciación a gran escala. Este hecho se ha visto reflejado en la generación de un enorme volumen de datos y en la creciente complejidad de las relaciones entre estos. En este sentido, las tecnologías de la información y comunicaciones han surgido como una herramienta imprescindible para posibilitar el almacenamiento, acceso e interpretación de estos datos de una forma eficiente. De esta forma, en los últimos años hemos sido testigos de una notable expansión de las aplicaciones computacionales en esta área: se han creado numerosas bases de datos que albergan información biológica de muy diversas fuentes, así como una extensa variedad de herramientas de análisis, tanto en el campo del software como del hardware. Las bases de datos moleculares han posibilitado, en estos últimos años, el estudio de la información genética con el desarrollo de programas que aceleran el análisis de numerosas variables. Así por ejemplo, puede leerse toda la secuencia de moléculas que compone el genoma de un organismo, es decir, el 'manual de instrucciones' que rige la formación del mismo y es posible la observación de su comportamiento dinámico bajo diversas condiciones experimentales. Ha surgido por lo tanto, una nueva disciplina denominada **Bioinformática** que agrupa conceptos tanto de la biología como de la informática, matemática o de la física y que permite cubrir la necesidad de manejo de las ingentes cantidades de información provenientes, tanto de la secuenciación de las macromoléculas (ADN, proteínas y glúcidos) como de las técnicas de análisis masivo del comportamiento de genes y proteínas. Consideramos que un estudiante que esté desarrollando su proyecto de Tesis Doctoral dentro de las áreas de la Genética, Biología Molecular e incluso Ecología Molecular o Zoología ha de enfrentarse en un momento u otro con un ordenador para realizar consultas a bases de datos o analizar secuencias de DNA o proteicas. Sin embargo, los resultados proporcionados por las diferentes herramientas informáticas no son

siempre fáciles de interpretar, tanto más cuanto que se tiende a ser autodidacta al afrontar este tipo de utilidades. El curso de Bioinformática planteado pretende llenar ese vacío que existe en la actualidad tal y como se comentó anteriormente.

El principal objetivo del curso que se plantea es el hacer llegar al alumnado el concepto de ordenador como una herramienta imprescindible en el laboratorio. A través de la bioinformática, se pretende mostrar hasta qué punto el empleo de los ordenadores puede ser útil en sus investigaciones. Como objetivos adicionales de pretende: (1) familiarizar al alumno con el término Bioinformática y mostrar las principales herramientas empleadas, sus posibilidades y las limitaciones de la bioinformática. (2) Describir las principales aplicaciones de la Bioinformática en la Genética e Ingeniería Genética, haciendo especial énfasis en los criterios que deben manejarse para ser críticos con los resultados que proporciona un programa informático. (3) Presentar los principios básicos en los que se basan los análisis bioinformáticos, esto es, la descripción del algoritmo que emplean dichos programas. Sin embargo, el objetivo no es que el alumno aprenda a desarrollar algoritmos o programación sino el conocimiento de las bases por las que se obtiene un resultado, con la finalidad de poderlo interpretar y evaluar de forma crítica. De esta forma la orientación final del curso es la formación de un biólogo como un usuario eficiente de los recursos bioinformáticos disponibles de tal forma que emplee las herramientas informáticas como una técnica más de laboratorio.

En cuanto al número de créditos y metodología de enseñanza, consideramos adecuado ofertarlo como un curso de 4 créditos de los cuales 2 se dedicarán a clases esencialmente magistrales en las que se explicarían todos los conceptos necesarios para familiarizarse con la bioinformática que serán detallados más adelante. Los 2 créditos restantes corresponderán a ejercicios prácticos, en los que se pondrían en práctica los conocimientos previamente adquiridos. El sistema de evaluación y calificación del alumno estará basado en la resolución de los ejercicios propuestos durante las sesiones de prácticas.

Nº DE CREDITOS ECTS: 5

TIPO: Obligatorio

SECUENCIA: 2º trimestre

CARÁCTER: Teórica / práctica

DESARROLLO: Presencial

BLOQUES TEMATICOS

Tema 1. Introducción a la Bioinformática: La información en el DNA y proteínas. Necesidad de herramientas computacionales en Biología. Servicios de información en Biología.

Tema 2.- Bases de datos en Biología. Tipos de bases de datos en biología. Bases de datos primarias: secuencias de DNA, secuencias de proteínas, estructuras de proteínas. Formatos de GenBank y EMBL. Bases de datos compuestas, derivadas o curadas. Bases de datos secundarias Bases de datos de clasificación estructural. Bases de datos para almacenar recursos genómicos

Tema 3.- Interrogación de Bases de datos. Interrogación booleana de las bases de datos. Búsquedas por texto. Búsquedas por similitud de secuencia. Formatos

de ficheros. Portales Entrez, SRS y SFGate. Estrategias de consulta. Bases de datos primarias de proteínas: Interrogación por el portal Expasy.

Tema 4.- Análisis de secuencias. Origen evolutivo de proteínas: conceptos de homología, analogía, paralogía y ortología. Bases de datos de ortólogos (COGs). Comparación y alineamientos de secuencias. Elaboración de alineamientos múltiples: Búsqueda de patrones y motivos funcionales a partir de un alineamiento múltiple: Clustal X. WUBLAST y PILEUP. Utilización de parámetros; bases de datos de alineamientos múltiples (Pfam y Blocks); bases de datos de motivos y dominios funcionales (Prosite y Domain).

Tema 5.- Análisis Filogenético. Principios de Sistemática Molecular: métodos de distancia, métodos de parsimonia. Estrategias en la búsqueda de árboles de máxima parsimonia: longitud del árbol, cálculos de los índices de consistencia y de homoplasia, arboles con y sin raíz, técnicas de remuestreo (“bootstrapping”), implicaciones de la Sistemática Molecular en la taxonomía. Utilización del programa PAUP4.0 sobre ejemplos reales para aplicación de métodos de distancia y de parsimonia y construcción de árboles. Comparación entre dichos métodos. Valoración de hipótesis mediante Bootstrap.

Tema 6.- Recursos bioinformáticos en el laboratorio. Análisis de secuencias individuales. Búsqueda de señales (promotores, repeticiones). Búsqueda de marcos abiertos de lectura. Predicción de genes. Diseño de cebadores específicos y degenerados a partir de datos de secuencia.

Tema 7. Análisis de genomas. Genómica comparada Proyectos secuenciación de genomas: de los mapas físicos a mapas de secuencia. Detección de regiones estructurales en un genoma. Genomas procariotas y eucariotas. Detección de un origen de replicación bacteriano: GC skew. Búsqueda de orfs y anotación de genomas. Sintenia. Caso práctico del organismo al mapa físico: análisis del genoma de *Buchnera aphidicola*. Genómica comparada: genoma humano (Venter, Watson y genoma de paciente con cáncer. Proyecto 1000 genomes).

Tema 8. Análisis masivo de genes: transcriptómica. Introducción a la tecnología de microarrays: oligonucleótidos, cDNA, productos de PCR. Medida de la expresión génica. Principales valores obtenidos de un experimento, interpretación y uso. Fuentes de error en análisis por microarrays. Identificación de las fuentes de error: errores sistemáticos y errores aleatorios: desplazamiento del Ratio (*Ratio shift*), intercambio de fluorocromos (*Dye-Swap*), contraste, efectos espaciales en ratio y contraste, problemas de adquisición de datos (cuantización y saturación). Errores aleatorios. Minimización del efecto de los errores: la plataforma de trabajo (PreP): acceso de datos, métodos analíticos, visualización, procesamiento de datos. Agrupamiento o aprendizaje no supervisado (*Clustering*): definiciones, casos y variables, medidas de distancia y de semejanza, métodos de aglomeración (Jerárquico, *k-means*, SOM, etc.). Introducción a la clasificación o aprendizaje supervisado. Contenidos Prácticos. Minimización de las fuentes de error. Pre-proceso de datos: filtrado, relleno de valores. Expresión diferencial. Agrupamiento (*Clustering*). Clasificación (selección de variables)

Tema 9. Bioinformática estructural. Niveles de estructura en proteínas. Predicción de estructura 2D y 3D. Búsqueda de modelos por homología y Threading (utilización del servidor Swiss-Model). Predicción de otras características estructurales: hélices transmembrana, motivos coiled-coil, accesibilidad al solvente e hidrofobicidad. Bases de datos de estructuras (PDB) y de clasificación jerárquica de estructuras (SCOP, CATH). Visualización de estructuras protéicas. Evolución y comparación de estructuras 3D. Estructura cuaternaria: interacción de proteínas. Diseño de fármacos. Predicción de estructuras secundarias de ARN (tRNAs y microRNAs). Herramientas de análisis de secuencias de ARN (paquete Vienna RNA y Mfold). Bases de datos de microRNAs.

Tema 10. Programación en Perl. Ejecución de scripts

Clases prácticas

Análisis de secuencias y análisis filogenético

Interrogación de Bases de datos.

Análisis masivo de genes: transcriptómica

BIBLIOGRAFIA BASICA

- Zvelebil, M. y T. Robins (2008) Understanding Bioinformatics. Garland Science
- Lesk, A. M. (2005) Introduction to Bioinformatics. Oxford University Press.
- David W. Mount (2004) Bioinformatics: sequence and genome analysis. New York : Cold Spring Harbor.
- Michael Barnes, Ian C Gray (2003) Bioinformatics for Geneticists, John Wiley & Sons