VNIVERSITAT DE VALÈNCIA

Facultat de Ciències Biològiques

GUÍA DOCENTE DE LA ASIGNATURA: BIOINFORMÁTICA
CURS 2008-09

I.- DATOS INICIALES DE IDENTIFICACIÓN

Nombre de la asignatura:	BIOINFORMÁTICA
Créditos ECTS	3
Carácter:	Obligatoria / Optativa
Posgrado:	Biotecnología / Biodiversidad
Máster:	Biología Molecular, Celular y Genética / Biodiversidad: Conservación y Evolución
Departamento:	Genética
Profesores responsables:	Francisco J. Silva correo electrónico: francisco.silva@uv.es Tlf: 9635 (43650) David Martínez correo electrónico: david.martinez@uv.es Tlf: 9635 (43644)

II.- INTRODUCCIÓN A LA ASIGNATURA

La materia Bioinformática se encuentra situada en el segundo curso del Máster en Biología Molecular, Celular y Genética. Esta asignatura es obligatoria y deberá ser cursada por todos los estudiantes independientemente de la especialidad elegida.

La asignatura es fundamentalmente práctica y por dicha razón los conocimientos teóricos serán impartidos simultáneamente con los prácticos en el aula de informática.

Originalmente la bioinformática fue definida como una materia interdisciplinar que incluía los campos de la biología, la informática, las

matemáticas y la estadística y cuyo objetivo era analizar los datos de secuencias biológicas, los contenidos y estructuras de los genomas, y la predicción y función de las proteínas. Con la llegada de la era de los genomas, la bioinformática ha extendido su campo de estudio al análisis de multitud de datos biológicos, entre ellos los derivados de los seres humanos y por tanto tiene actualmente una gran importancia en la investigación biomédica.

Los 3 créditos ECTS asignados a esta materia hacen que los conocimientos y habilidades que se adquirirán en este curso puedan ser definidos como una introducción a la bioinformática. A partir de ellos los estudiantes podrán continuar su aprendizaje, bien de forma autodidacta o, en caso de haber elegido la especialidad de Genética, ampliando sus conocimientos sobre herramientas bioinformáticas para el análisis de genomas en la asignatura de Genómica comparada.

III.- VOLUMEN DE TRABAJO

TRABAJO PRESENCIAL

Asistencia a sesiones teórico-prácticas:
2 horas/semana x 9 semanas = **18 horas.**

Asistencia a tutorías personales:
promedio de **1 hora.**

Asistencia a pruebas de evaluación:
hasta un total de **2,5 horas.**

Realización de una encuesta sobre la asignatura:
0,5 horas.

TOTAL PRESENCIAL= 22 HORAS

TRABAJO NO PRESENCIAL

Preparación previa de sesiones teórico-prácticas:
0,5 horas x 10 sesiones = **5 horas.**

Estudio de contenidos teórico-prácticos:
hasta un total de **20 horas**

Ejercicios y trabajos prácticos realizados en grupo:
Resolución y análisis de diversos ejercicios sobre la materia
hasta un total de **22 horas**

Elaboración de un informe sobre su realización:
6 horas.

TOTAL NO PRESENCIAL= 53 HORAS

Tabla resumen del volumen de trabajo:

	Horas/curso
Asistencia a sesiones teórico-prácticas	**18**
Asistencia a tutorías personales	**1**
Asistencia a pruebas de evaluación	**2,5**
Realización de una encuesta sobre la asignatura	**0,5**
Preparación previa de sesiones teórico-prácticas	**5**
Estudio de contenidos teórico-prácticos	**20**
Ejercicios y trabajos prácticos realizados en grup	**28**
TOTAL VOLUMEN DE TRABAJO	**75**

IV.- OBJETIVOS GENERALES

Adquirir los conocimientos teóricos y prácticos relativos a los siguientes puntos:

(1) Gestión de proyectos de secuenciación
(2) Extracción de información o secuencias de las bases de datos moleculares
(3) Búsqueda de secuencias en bases de datos por similitud
(4) Alineamiento de secuencias de DNA y proteínas

V.- CONTENIDOS

Conocer cómo obtener secuencias definitivas de calidad a partir de los datos proporcionados por un secuenciador.

Conocer como preparar las secuencias proporcionadas por un secuenciador para su posterior análisis.

Conocer los distintos métodos para el editado de las secuencias procedentes de un secuenciador.

Conocer la existencia y modo de acceso a las bases de datos de secuencias generales

Conocer como las secuencias se organizan en las bases de datos

Conocer los formatos de los archivos de secuencia

Conocer las bases de datos de dominios de proteínas y de familias de RNAs no codificantes.

Conocer la existencia y modo de acceso a las bases de datos genómicas MBGD, Microbesonline y Ensembl

Conocer los métodos y algoritmos que se utilizan para la búsqueda en bases de datos por similitud, especialmente el algoritmo BLAST.

Saber determinar cuál es el tipo de BLAST más adecuado en cada caso.

Conocer los métodos de alineamiento de secuencias.

Entender el significado de las matrices de sustitución aminoacídica.

Conocer los métodos para realizar alineamientos locales

VI.- DESTREZAS A ADQUIRIR.

Extraer una secuencia a partir de un fichero procedente de un secuenciador.

Detectar y ocultar secuencias de vector y secuencias contaminantes.

Ensamblar secuencias.

Visualizar y editar manualmente los “contigs”.

Introducir anotaciones.

Detectar solapamientos en segunda instancia.

Eliminar secuencias de una base de datos.

Obtener información de las bases de datos mediante sistemas de búsqueda por palabras claves.

Extraer simultáneamente secuencias de genes o proteínas homólogas de varias especies en la base de datos MBGD

Observar de forma comparada segmentos de genomas que contengan genes homólogos.

Comparar las capacidades metabólicas y funcionales derivadas de los genomas de dos especies.

Realizar árboles filogenéticos con las herramientas de las bases de datos MBGD y Microbesonline.

Utilizar el algoritmo BLAST y comprender el significado de su resultado.

Ser capaz de seleccionar las opciones más adecuadas para cada tipo de análisis BLAST (tipo de BLAST, restricción por especie, restricción de secuencia, umbrales de *expected value*).

Ser capaz de utilizar las mejores opciones del método CLUSTAL para alinear secuencias de proteínas, secuencias de genes codificantes, secuencias de DNA no codificantes o pseudogenes.

Ser capaz de utilizar las mejores opciones para alinear DNA genómico y cDNAs.

Ser capaz de editar y modificar manualmente segmentos incorrectos de un alineamiento.

Ser capaz de realizar un árbol filogenético con el algoritmo del Neighbor joining con el programa MEGA.

VII.- HABILIDADES SOCIALES

Capacidad para trabajar, resolver problemas y realizar ejercicios en grupo.

Capacidad para el trabajo *online* con programas de Internet.

Dominio de una lengua extranjera, en especial en lo que respecta a la comprensión de textos sencillos de carácter científico en inglés.

Capacidad de análisis y de síntesis.

Capacidad de construir un texto escrito comprensible y organizado.

VIII.- TEMARI I PLANIFICACIÓ TEMPORAL

La bioinformática es una asignatura eminentemente práctica en la que es necesario adquirir ciertos conocimientos teóricos sobre los métodos y algoritmos en los que se basan los diversos programas. Se impartirán 8 clases teórico-prácticas de 2 horas. Durante las clases se alternarán las explicaciones teóricas que centren los problemas a resolver y los métodos conocidos para hacerlo, con el uso de los diversos programas y/o páginas web donde se pueden ejecutar.

A continuación de las clases presenciales los estudiantes continuarán trabajando en la resolución de diversas cuestiones planteadas por los profesores. Durante o al final del periodo de clases los estudiantes entregarán un informe sobre los ejercicios o problemas planteados para su resolución.

TEMARIO DE LAS SESIONES TEORICO-PRÁCTICAS

Las sesiones descritas a continuación están pensadas para ser realizadas durante dos horas, con una periodicidad aproximada de una sesión por semana.

Sesión 1. Introducción al paquete *Staden*.

Presentación del paquete y de sus funciones. Obtención e introducción de ficheros procedentes de un secuenciador. Obtención de secuencias de vectores. Obtención e introducción de otro tipo de secuencias.

Sesión 2. Utilización del programa *Pregap4*.

Preparación de las secuencias. Extracción de la parte útil de una secuencia. Conversión de formatos. Ocultación de partes de una secuencia con mala calidad. Detección y ocultación de secuencias de vector. Búsqueda y eliminación de secuencias contaminantes. Personalización de módulos con tareas automatizadas.

Sesión 3. Utilización del programa *Gap4*.

Ensamblado inicial de secuencias. Visualización y edición manual de “contigs”. Visualización y recuperación de secuencias ocultas. Introducción de anotaciones. Criterios para la edición rápida de contigs. Obtención de secuencias consenso.

Sesión 4. Ensamblado avanzado de secuencias.

Automatización del proceso de ensamblado. Búsqueda de solapamientos en segunda instancia. Introducción de nuevas secuencias a una base de datos preexistente. Eliminación de secuencias de una base de datos.

Sesión 5. Bases de datos moleculares I.

Bases de datos generales (GenBank, EBI y DDBJ). Acceso a secuencias de DNA genómico, ESTs y proteínas. Sistemas de búsquedas por palabras

clave (Entrez, SRS, EntrezGene, UniGene, Expasy). Formato de las secuencias.

Sesión 6. Bases de datos moleculares II.

Bases de datos de Pfam y Rfam. Envio de secuencias a las bases de datos. Bases de datos de genomas (MBGD, Microbesonline, Ensembl)

Sesión 7. Búsqueda por similitud de secuencia.

Alineamientos locales y globales Algoritmos de búsqueda por similitud. El algoritmo BLAST. Tipos de BLAST.

Sesión 8. Alineamientos de múltiples secuencias.

Alineamiento de proteínas. Matrices de sustituciones aminoacídicas. Alineamiento de secuencias nucleotídicas. Alineamiento de genes codificantes de proteínas, pseudogenes, DNA genómico y ESTs.

Sesión 9. Realización de ejercicios tutorizados sobre las sesiones 1 a 8.

IX.- BIBLIOGRAFÍA DE REFERENCIA

a) Bibliografía básica:

Analysing Sequences Using the Staden Package and EMBOSS. Introduction to Bioinformatics. A Theoretical and Practical Approach por Staden, R., Judge, D. P. and Bonfield, J. K. (2003). Eds. Stephen A. Krawetz and David D. Womble. Human Press Inc., Totawa, NJ 07512

Bioinformatics: Sequence and Genome Analysis, Second Edition por David Mount (2004) publicado por Cold Spring Harbor Press. Un libro amplio y bien documentado sobre bioinformática

Bioinformatics and Functional Genomics, por Jonathan Pevsner (2003) publicado por Wiley_Liss. Una introducción a la bioinformática y la genómica con menos detalles que Mount pero más fácil de seguir y de entender los conceptos. Incluye muchos ejercicios prácticos y direcciones *web*.

Introducción a la Bioinformática, por T.K. Attwood y D. J. Parry-Smith (2002). publicado por Prentice Hall. Traducción al castellano de la edición en inglés del año 1999. Menos amplio y algo más atrasado que los anteriores pero con la facilidad idiomática.

b) **Páginas web** con recursos bioinformáticos

- La base de datos del European Molecular Biology Laboratory (EMBL) del European Bioinformatics Institut (EBI, http://www.ebi.ac.uk/index.html) situado en Hixton (Inglaterra)
- La base de datos GenBank del National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) de National Institutes of Health situado en Bethesda (USA)

- La base de datos DNA Datababe of Japan (DDBJ, http://www.ddbj.nig.ac.jp) situada en Mishima Japón.
- SRS (Sequence retrieval system, http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+top). Sistema de búsqueda por palabras clave del EMBL
- Entrez (http://www3.ncbi.nlm.nih.gov/Entrez/) Sistema de búsqueda por palabras clave del con similares características del NCBI.
- ExPASy (Expert Protein Analysis System, http://www.expasy.ch/). Servidor de proteómica del Swiss Institute of Bioinformatics (SIB). está dedicado a la biología molecular con un énfasis especial en los datos de proteínas.
- EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) Sistema de búsqueda en una base de datos cuyas entradas son genes.
- UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) Sistema de búsqueda sobre una base de datos no redundante donde las secuencias correspondientes al mismo gene aparecen agrupadas en una única entrada.
- MBGD (Microbial genome database for comparative analysis, http://mbgd.genome.ad.jp/) es una base de datos del Okazaki National Research Institutes para el análisis comparado de secuencias de genomas completos microbianos. El objetivo principal es la identificación de genes ortólogos, parálogos, orden génico, función, etc.
- MicrobesOnline (http://www.microbesonline.org) está diseñada específicamente para facilitar los estudios comparados entre genomas procariotas. Incluye la predicción de operones y regulones. Permite la comparación de los mapas genómicos, y análisis metabólicos con las rutas metabólicas de KEGG.
- Ensembl (http://www.ensembl.org/index.html) es un proyecto conjunto del EMBL-EBI (Laboratorio Europeo de Biología Molecular - Instituto Europeo de Bioinformática) y del Sanger Institute cuyo objetivo es el desarrollo y mantenimiento de un sistema informático para el trabajo con un conjunto de genomas eucariotas seleccionados entre los que se encuentra el hombre y los primates.
- Pfam (Protein families Database of alignments and HMMs, http://www.sanger.ac.uk/Pfam/) es una gran colección de alineamientos de secuencias y de "modelos de Markov escondidos (HMMs)" que cubre muchos de los dominios más frecuentes de proteínas. Permite no sólo ver la estructura en dominios de muchas proteínas, sino que le enviemos una nueva secuencia y mediante un programa nos muestre cual sería su estructura de dominios.
- Base de datos Rfam (http://www.sanger.ac.uk/Software/Rfam/). Contiene familias de RNAs no codificantes.

Direcciones para realizar blast:

- GenomeNet (Centro de Bioinformática de Univ. Kyoto) http://blast.genome.jp/)
- NCBI (http://www.ncbi.nlm.nih.gov/BLAST/)
- EMBO (http://www.ebi.ac.uk/Tools/homology.html)
- KEGG (http://www.genome.ad.jp/kegg-bin/SearchGenes?blast)

  (http://www.genome.ad.jp/kegg-bin/SearchGenes?blast+genome)

- MEGA es un programa que permite el alineamiento de secuencias, su análisis y la elaboración de árboles filogenéticos (http://www.megasoftware.net/).
- MEME (http://meme.sdsc.edu/meme/intro.html) es una herramienta para descubrir motivos en un grupo de secuencias de DNA o proteínas relacionadas.
- RSA-tools - convert-matrix (http://rsat.ulb.ac.be/rsat/convert-matrix_form.cgi) crea una matriz de perfil de posiciones (PSSM) a partir de un alineamiento.
- WebLogo (http://weblogo.berkeley.edu/) Genera Logos de secuencia a partir de alineamientos.
- MFOLD (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html) predice estructuras secundarias en el RNA
- Eukaryotic GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi) predice los límites de los genes en eucariotas y por tanto exones e intrones.
- Softberry (http://www.softberry.com/berry.phtml) es una compañía que desarrolla programas bioinformáticas tiene muchos programas para la detección de genes tanto en eucariotas como en procariotas, para localizar promotores, etc.

X.- METODOLOGÍA

El desarrollo de la asignatura se estructura en trabajos presenciales y no presenciales:

Trabajo presencial:

A) Ocho sesiones de clases en aula de informática sobre diversas bases de datos y como obtener información de ellas. También se enseñará a los estudiantes el uso de diversos programas informáticos que en algunos casos correrán localmente y en otros será necesario realizar los análisis *on line*.
B) Tutorías
C) Examen
D) Una encuesta sobre la asignatura donde poder localizar las partes mejorables o los contenidos no incluidos y que puedan ser de interés para los estudiantes.

Trabajo no presencial:

A) Ejercicios prácticos y problemas bioinformáticos que deberán ser resueltos por los estudiantes en grupos reducidos haciendo uso del aula de informática. Los resultados de algunos de estos ejercicios serán guardados en la sección de tareas del aula virtual. Para algunos de ellos se pedirá un informe escrito que se presentará al final de las clases.
B) Estudio de los contenidos y preparación previa de las clases. Se indicará a los estudiantes los capítulos de los libros recomendados donde pueden leer antes de la clase su contenido u otros materiales equivalentes.

XI.- EVALUACIÓN DEL APRENDIZAJE

La evaluación del aprendizaje de los estudiantes en **primera convocatoria** se realizará mediante la valoración de los siguientes apartados:

1) Un **examen** teórico-práctico que se realizará en el aula de informática y que podrá constar tanto de preguntas sobre los conocimientos de teoría como de ejercicios que se deberán resolver utilizando los programas estudiados en el curso. Esta prueba permitirá obtener hasta **7 puntos** y se realizará tras la finalización de las clases.

2) La resolución de los **ejercicios** y problemas bioinformáticos comentados en el apartado X (Metodología). Este apartado valdrá hasta **3 puntos**. Si se obtiene una puntuación igual o superior a 1,5 se podrá guardar la misma para la segunda convocatoria.

3) Finalmente el estudiante dispondrá de un portafolio donde se irán acumulando puntos asociados a la valoración que el profesor realice sobre su interés en la asignatura expresado como su participación en las clases, las contestaciones a las preguntas que realice el profesor durante las sesiones presenciales, su asistencia a tutorías personales y/o cualquier otro tipo de actividad llevada a cabo por el estudiante en relación con la asignatura. Se podrá conseguir hasta **1 punto extra** en la calificación final de la asignatura.

Para superar la asignatura en primera convocatoria será necesario obtener **5 puntos** de calificación total y un mínimo de **2,5 puntos** en el examen y de **1 punto** en los ejercicios.

Para superar la asignatura **en segunda convocatoria**, habrá que superar un único examen teórico-práctico similar al planteado en el apartado XI-1. En el caso que el estudiante hubiese obtenido una puntuación igual o superior a 1,5 en la resolución de los ejercicios (ver apartado XI-2) estos puntos se sumaran a los del examen, que valdrá hasta **7 puntos**. En el caso de que la puntuación de los ejercicios del apartado XI-2 fuese inferior a 1,5 o fuese inexistente, el valor del examen pasará a ser de hasta 10 puntos