PROGRAMA de BIOINFORMÁTICA

Carrera/s: *Licenciatura en Biotecnología*

Asignatura: *Bioinformática*

Núcleo al que pertenece: *Obligatorio (Ciclo Superior)*[1]

Profesores/as: *Carolina S. Cerrudo, Néstor G. Iglesias, Solange AB Miele, Nicolás Paloppoli.*

Correlatividades previas: *Genética Molecular*

Objetivos:

Que las/os estudiantes desarrollen las capacidades cognitivas y analíticas necesarias para utilizar adecuadamente las herramientas básicas de bioinformática en el análisis, interpretación y uso de la información molecular de origen biológico.

Contenidos mínimos:

Niveles de información biológica. Acceso remoto a bancos de datos, algoritmos de búsqueda. Bancos de datos genéticos. Análisis de secuencias biológicas. Identidades y similitudes secuenciales y estructurales. Minería de datos (data mining): búsqueda de patrones y motivos. Teoría de la información y su aplicación al estudio de las secuencias biológicas. Aspectos composicionales en ácidos nucleicos y proteínas. Evolución molecular: filogenia y mecanismos de transferencia de material genético. Micro y Macroevolución. Predicción de la estructura secundaria en ácidos nucleicos. Predicción de la estructura secundaria en proteínas. Aproximaciones a la predicción de estructura terciaria en proteínas: modelado por homología (homology modeling). Metodologías relacionadas con proteómica. Métodos ómicos para la caracterización de la materia viva.

Carga horaria semanal: 6 Hs

Programa analítico:

Unidad 1: Bioinformática, consideraciones generales. Vías de acceso a la

[1] En plan vigente, Res CS N° 125/19. Para el plan Res CS N° 277/11, pertenece al Núcleo Básico. Para el Plan Res CS N° 181/03 pertenece al Núcleo Orientado.

información según la problemática. Bases de datos: características, acceso y principales herramientas para la búsqueda y el análisis de genes. Los proyectos genoma y el problema del análisis e interpretación de secuencias nucleotídicas y aminoacídicas.

Unidad 2: Estrategias básicas para la búsqueda de similitud entre dos o más secuencias, nucleotídicas o aminoacídicas. Principales algoritmos: métodos basados en matrices de puntos (Dot Plot, Dotlet, etc.), métodos basados en un análisis global (Clustal), métodos basados en un análisis local (BLAST). Ensamble de secuencias derivadas de secuenciaciones automáticas (CONTIGS).

Unidad 3: Búsqueda de patrones y motivos en secuencias nucleotídicas y aminoacídicas. Principales algoritmos. Bases de datos de motivos (PROSITE, Pfam, etc.). Análisis basado en perfiles de búsqueda. Estimaciones probabilísticas y validez de los resultados. Una aproximación a la predicción de posibles funciones biológicas de una secuencia nucleotídica o aminoacídica.

Unidad 4: La teoría de la información: conceptos básicos. La teoría de la información y el análisis de secuencias. Complejidad informativa (entropía) global y local. Análisis de secuencias basado en aspectos informativos. Herramientas para el análisis individual y múltiple. Aplicaciones prácticas: logos de secuencias, diseño de primers basado en complejidad informativa.

Unidad 5: Análisis de secuencias basado en aspectos composicionales. Abundancia relativa de oligonucleótidos cortos (*GENOMIC SIGNATURE*). Frecuencias nucleotídicas y aminoacídicas. Uso de codones.

Unidad 6: La filogenia basada en secuencias nucleotídicas o aminoacídicas. El problema del análisis filogenético basado en segmentos de secuencias. PHYLIP (*Phylogeny Inference Package*), criterios para el análisis e interpretación de resultados. La filogenia y el criterio de *TOTAL EVIDENCE*. El problema de la transferencia horizontal y la recombinación. T-REX, SIMPLOT y otros.

Unidad 7: La predicción de estructuras secundarias en ácidos nucleicos. Principales criterios y algoritmos (FOLD, MULFOLD, RNADraw, RNAStructure). La predicción de estructuras óptimas y subóptimas. Análisis comparativo de patrones de plegamiento entre las formas óptimas y subóptimas. Validez de los resultados.

Unidad 8: El análisis de secuencias y la predicción de estructuras secundarias en proteínas. Análisis y predicción de propiedades fisicoquímicas relacionadas con la estructura. Principales criterios y algoritmos (hidropatía, anfipaticidad,

etc.). Estructuras secundarias (GOR IV, SCOP, CATH, JPred). Validez de los resultados.

Unidad 9: Genómica funcional: proteomas, transcriptomas, regulomas. Generación de datos. Bases de datos y herramientas de análisis.

Unidad 10: Diseño de oligonucleótidos. Variantes. Criterios generales y alternativos. Estrategias de diseño.

Unidad 11: Tecnologías NGS (*Next Generation Sequencing*). Variantes tecnológicas. Estrategias de ensamblado de contigs. Criterios de calidad. Otras tecnologías ómicas para el estudio de macromoléculas biológicas y sus niveles de información. Aplicaciones.

Trabajos prácticos de laboratorio

TP N°1: Acceso al GenBank y sus bases de datos. Acceso a las principales bases de datos. Búsqueda de genes, secuencias, estructuras y funciones.

TP N°2a: Alineamientos mediante matriz de puntos, *Dot Plot*. Estrategias básicas para la búsqueda de similitud entre secuencias nucleotídicas o aminoacídicas. Métodos basados en matrices de puntos (*Dot Plot*). Análisis de ejemplos específicos en programas locales y servidores internacionales.

TP N°2b: Blast. Métodos basados en un análisis local (BLAST). Análisis de ejemplos específicos en servidores locales e internacionales. Estrategias básicas para la búsqueda de similitud entre dos secuencias nucleotídicas o aminoacídicas y búsquedas en la base de datos del NCBI.

TP N°2c: *Dynamic Programming Methods* (DPM) y Clustal X. Estrategias básicas para la búsqueda de similitud entre dos o más secuencias. Métodos basados en programación dinámica (algoritmos Needleman-Wunsch y Smith-Waterman) y métodos basados en un análisis global (Clustal). Análisis de ejemplos específicos en programas locales y servidores internacionales.

TP N°3: Búsqueda de Motivos y Patrones. Búsqueda de patrones y motivos en secuencias nucleotídicas y aminoacídicas. Principales algoritmos. Bases de datos de motivos (PROSITE, Pfam, CD-Search). Estrategias de mostración de resultados. Análisis de ejemplos específicos en Prosite (http://prosite.expasy.org/scanprosite/), Pfam (http://pfam.xfam.org/), Interproscan (https://www.ebi.ac.uk/interpro/search/sequence-search), CD-Search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

TP N°4a: Complejidad informativa. La teoría de la información y el análisis de secuencias. Complejidad informativa (entropía) global y local. Análisis de ejemplos específicos con programas locales. Estrategias de mostración de resultados.

TP N°4b: *Sequence Logos*. La teoría de la información y el análisis de secuencias. Herramientas para el análisis individual y múltiple. Aplicaciones prácticas: logos de secuencias. Análisis de ejemplos específicos en servidores internacionales. Estrategias de mostración de resultados.

TP N°5: Aspectos composicionales. Análisis de secuencias nucleotídicas y aminoacídicas basado en aspectos composicionales. Frecuencias de residuos (nucleótidos o aminoácidos), distintos tipos de aproximaciones. Abundancia relativa de oligonucleótidos cortos (*Genomic Signature*). Abundancia vs Frecuencia. Asimetría composicional en genomas, alternativas de análisis. Análisis de ejemplos específicos en servidores internacionales, GenSkewApp.jar (http://genskew.csb.univie.ac.at/), CG view (http://stothard.afns.ualberta.ca/cgview_server/). Estrategias de mostración de resultados.

TP N°6a: MEGA. La filogenia basada en secuencias nucleotídicas o aminoacídicas. Principales programas relacionados con la inferencia filogenética (MEGA, PHYLIP); análisis por distancia y por parsimonia; criterios para el análisis e interpretación de resultados. Estrategias de mostración de resultados.

TP N°6b: Reticulogramas. La filogenia basada en secuencias nucleotídicas o aminoacídicas. El problema del análisis filogenético basado en segmentos de secuencias. Evolución molecular: filogenia y mecanismos de transferencia de material genético. Detección de secuencias recombinantes utilizando T-Rex; criterios para el análisis e interpretación de resultados.

TP N°6c: SimPlot. La filogenia basada en secuencias nucleotídicas o aminoacídicas. El problema del análisis filogenético basado en segmentos de secuencias. Evolución molecular: filogenia y mecanismos de transferencia de material genético. Detección de secuencias recombinantes y caracterización de parentales utilizando Simplot; criterios para el análisis e interpretación de resultados.

TP N°7a: Estructuras secundarias en ácidos nucleicos. La predicción de estructuras secundarias en ácidos nucleicos. Principales criterios y algoritmos (FOLD, MULFOLD, RNADraw). Predicción de estructuras óptimas y subóptimas utilizando el programa **RNAdraw.exe**, y el servidor de Michael Zuker (http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form). Análisis comparativo de patrones de plegamiento entre las formas óptimas y subóptimas. Validez de los resultados. Estrategias de mostración de resultados.

TP N°7b: Estructuras secundarias en ácidos nucleicos. Predicción de estructuras óptimas y subóptimas utilizando el programa **RNAStructure** (**v 3.71**). Análisis comparativo de patrones de plegamiento entre las formas óptimas y subóptimas. Validez de los resultados. Estrategias de mostración de resultados. Comparación con los resultados obtenidos en el trabajo práctico anterior.

TP N°8a: Análisis de hidropatía en proteínas. Análisis y predicción de propiedades fisicoquímicas relacionadas con la estructura. Principales criterios y algoritmos (hidropatía, anfipaticidad, etc.). Predicción de hidrofobicidad utilizando el servidor Protscale (http://web.expasy.org/cgi-bin/protscale/protscale.pl), Estrategias de mostración de resultados.
TP N°8b: Análisis de proteínas. El análisis de secuencias, caracterización de estructura primaria de proteínas y la predicción de estructuras secundarias en proteínas. Caracterización de proteínas utilizando el servidor http://www.expasy.org/tools. Predicción de estructuras secundarias utilizando diferentes servidores internacionales (Jpred, GOR, PredictProtein). Validez de los resultados. Estrategias de mostración de resultados.

TP N°8c: Predicción de anfipaticidad en estructuras secundarias de proteínas. Análisis y predicción de propiedades fisicoquímicas relacionadas con la estructura. Principales criterios y algoritmos (hidropatía, anfipaticidad, etc.). Predicción de anfipaticidad en estructuras secundarias de proteínas utilizando el servidor Heliquest, http://heliquest.ipmc.cnrs.fr/cgi-bin/ComputParamsV2.py. Estrategias de mostración de resultados.

TP N°8d: Comparación de estructuras 3D. Aproximaciones a la predicción de estructura terciaria en proteínas (*homology modelling*, *folding recognition*, etc.). Alineamientos de estructuras terciarias de proteínas utilizando los servidores Dali server (http://ekhidna.biocenter.helsinki.fi/dali_server/) y **PDBeFold** (http://www.ebi.ac.uk/msd-srv/ssm/). Estrategias de mostración de resultados.

TP N°9a: Análisis de Redes Biológicas. Genómica funcional: genomas, proteomas, transcriptomas, regulomas, etc. Bases de datos y herramientas de análisis. Metodologías adicionales relacionadas con la genómica funcional. Generación y análisis de redes biológicas utilizando el servidor ***GeneMANIA*** (http://www.genemania.org/). Estrategias de mostración de resultados.

TP N°9b: Análisis de Redes Biológicas. Genómica funcional: genomas, proteomas, transcriptomas, regulomas, etc. Bases de datos y herramientas de análisis. Metodologías adicionales relacionadas con la genómica funcional. Generación y análisis de redes biológicas utilizando el servidor ***STRING*** (http://string-db.org). Estrategias de mostración de resultados. Comparación con los resultados obtenidos en el trabajo práctico anterior.

TP N°10: Diseño de *Primers* utilizando el Blast. En este Trabajo Práctico analizaremos la forma de diseñar *primers* utilizando el servidor Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) de la sección Specialized BLAST.

Bibliografía:

Sequence analysis primer. M. Gribskov and J. Deveraux. 1991. UWBC

Biotechnical Resource Series. Stockton Press. New York. USA.

Molecular evolution: Computer analysis of protein and nucleic acid sequences. R. F. Doolittle. 1990. Methods in Enzymology, volume 183. Academic Press.California. USA.

Computer Methods for Macromolecular Sequence Analysis. R. F. Doolittle. 1996. Methods in Enzymology, volume 266. Academic Press. California. USA.

Bioinformatics. Methods and Protocols. S. Misener and S.A. Krawetz. 1999. Humana Press. New Jersey. USA.

Computational Methods in Molecular Biology. Salzberg S.L., Searls D.B. and Kasif S. 1998. Elsevier Science. USA.

Theoretical and Computational Methods in Genome Research. Suhai. 1998. Kluwer Academic Publishers. USA.

Bioinformatics for Biologists. Pavel Pevzner, Ron Shamir. 2011 Cambridge University Press. UK.

Next-Generation DNA Sequencing Informatics. Stuart M. Brown. 2013. Cold Spring Harbor Laboratory. USA.

An Introduction to Bioinformatics Algorithms (Computational Molecular Biology) Neil C. Jones and Pavel A. Pevzner 2004 MIT Press. USA.

Biological Data Mining. Jake Y. Chen, Stefano Lonardi 2009 Chapman & Hall/CRC

Hidden Markov Models for Bioinformatics. Timo Koskinen. 2001. Kluwer Academic Publishers. USA.

Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins. Andreas D. Baxevanis, B. F. Francis Ouellette. 2004. Wiley, John & Sons, Inc. USA.

Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. Kulkarni P, Frommolt P. Comput Struct Biotechnol J. 2017 Oct 25;15:471-477.

Next-Generation DNA Sequencing Informatics. Stuart M. Brown. 2013. Cold Spring Harbor Laboratory. USA.

Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools. Quoc Nam Tran and Hamid Arabnia. 2015. Elsevier and Morgan Kaufmann. USA.

Algorithms for next-generation sequencing. Wing-Kin Sung. 2017. CRC Press. USA

Apuntes de la asignatura.

Publicaciones periódicas seleccionadas.

INTERNET.

La bibliografía que no se encuentra en la Biblioteca de la UNQ es suministrada por los docentes, ya sea porque se dispone de las versiones electrónicas y/o se dispone del ejemplar en el grupo de investigación asociado.

Organización de las clases:
A partir de situaciones problemáticas concretas, presentadas por la/os docentes, esta asignatura se desarrollará principalmente mediante seminarios teóricos y de discusión, con la activa participación de los estudiantes, el uso de las herramientas informáticas básicas, tanto en forma local como remota (INTERNET), y la exposición y discusión de *papers* seleccionados. Durante cada clase se destinará tiempo a clases teóricas y luego al desarrollo de trabajos prácticos que afiancen los contenidos teóricos abordados anteriormente, empleando para ello aulas equipadas con computadoras.

Modalidad de evaluación:
Desde el punto de vista de la evaluación, en esta asignatura se considera como un aspecto muy importante la activa participación en todas las instancias. El/la estudiante deberá asistir al menos al 75% de las clases. Además:

- *Habrá dos exámenes, uno escrito y uno en computadora*
- *Las exposiciones de papers serán evaluadas conceptualmente.*
- *Al término del curso deberá presentarse un trabajo final, sobre un tema asignado por la/os docentes, desarrollado en formato paper (Título, Introducción, Materiales y Métodos, Resultados, Discusión y Bibliografía).*

Aprobación de la asignatura según Régimen de Estudios vigente de la Universidad Nacional de Quilmes:
La aprobación de la materia bajo el régimen de regularidad requerirá: Una asistencia no inferior al 75 % en las clases presenciales previstas, y cumplir con al menos una de las siguientes posibilidades:

(a) la obtención de un promedio mínimo de 7 puntos en las instancias parciales de evaluación y de un mínimo de 6 puntos en cada una de ellas.
(b) la obtención de un mínimo de 4 puntos en cada instancia parcial de evaluación y en el examen integrador, el que será obligatorio en estos casos. Este examen se tomará dentro de los plazos del curso.

Los/as alumno/as que obtuvieron un mínimo de 4 puntos en cada una de las instancias parciales de evaluación y no hubieran aprobado el examen integrador mencionado en el Inc. b), deberán rendir un examen integrador, o en su reemplazo la estrategia de evaluación integradora final que el programa del curso establezca, que el cuerpo docente administrará en los lapsos estipulados por la UNQ.

Modalidad de evaluación exámenes libres:
En la modalidad de libre, se evaluarán los contenidos de la asignatura con un examen escrito, un examen oral e instancias de evaluación similares a las realizadas en la modalidad presencial. Los contenidos a evaluar serán los especificados anteriormente incluyendo demostraciones teóricas, laboratorios y problemas de aplicación.

Anexo II

CRONOGRAMA TENTATIVO BIOINFORMÁTICA

Semana	Tema/Unidad	Actividad*				Evaluación
		Teórico	Práctico			
			Resol. Proble-mas	Labora-torio	Otros (Exposiciòn de *papers*)	
1	Introducción - Unidad 1 y Unidad 2	X	X	X		
2	Unidad 2	X	X	X		
3	Unidad 3	X	X	X		
4	Unidad 4	X	X	X		
5	Unidad 4 y 5	X	X	X		
6	Unidad 5 y Práctico TP Final	X	X	X		
7	Consulta y Examen escrito	X				X
8	Unidad 6	X				
9	Unidad 6 y Presentación de Papers	X	X	X	X	
10	Práctico TP Final, Presentación de Papers y Unidad 7	X	X	X	X	
11	Unidad 7, Práctico TP Final y Presentación de Papers	X	X	X	X	
12	Unidad 8	X	X	X		
13	Unidad 8 y 9, Presentación de Papers	X	X	X	X	
14	Unidad 9 y 10	X	X	X		
15	Unidad 11 y Práctico TP Final	X	X	X		
16	Consulta y Examen en máquina	X				X
17	Consulta y Práctico TP Final	X	X	X		X
18	Recuperatorios y Fecha límite de entrega Trabajo Práctico Final					X
19	Examen Integrador y Cierre de actas					X

***INDIQUE CON UNA CRUZ LA MODALIDAD**

Nota: Las actividades de Resolución de problemas y Laboratorio coinciden ya que se los problemas se resuelven, con la guía de los docentes, en el laboratorio de computadoras.