Título del Curso: Herramientas Bioinformáticas para el análisis de estructura, desorden e interacciones de proteínas

Docente responsable: Dra. Lucía Beatriz Chemes, Prof. Adjunta UNSAM

Docentes auxiliares: Dra. Juliana Glavina (JTP UNSAM), Lic. Heli Magalí García Álvarez (Ay1 UNSAM), Lic. Alejandro Ricci (Ay1 UNSAM).

Docentes invitados: Dr. Toby Gibson (EMBL Heidelberg), Dr. Nicolás Palopoli (UNQ)

Tipo de curso:
Curso de posgrado (UNSAM y otras universidades) y materia optativa de grado (UNSAM).

Objetivos:

Este curso presentará un conjunto de herramientas bioinformáticas que permiten explorar la estructura, regiones funcionales e interacciones de proteínas que participan en las redes de señalización eucariotas a través de clases teóricas y sesiones prácticas intensivas que enseñarán a explorar la arquitectura modular de las proteínas, con el fin de poder predecir sus posibles funciones e interacciones. Los alumnos/as adquirirán conceptos básicos de estructura-función de proteínas y aprenderán a manejar un conjunto de bases de datos y predictores que permiten analizar dominios y regiones funcionales, estructura de proteínas y sus interacciones.

El curso proveerá una base teórica y práctica para el estudio de relaciones estructura-función de proteínas utilizando herramientas bioinformáticas. Se introducirán bases de datos y herramientas que permiten identificar dominios funcionales de proteínas, predecir la estructura de proteínas a partir de la secuencia, discernir regiones plegadas y flexibles de proteínas y analizar diferentes tipos de interacciones proteína-proteína. A lo largo de los trabajos prácticos, los alumnos adquirirán manejo de herramientas de visualización y análisis de estructura y secuencia de proteínas utilizando los programas Chimera y Jalview. El curso está dirigido a estudiantes avanzados de grado, estudiantes de doctorado y postdocs/investigadores con formación experimental que deseen adquirir conceptos de bioinformática estructural, y a alumnos con formación en bioinformática que no tengan conocimientos previos de estructura-función de proteínas. Los alumnos adquirirán competencia técnica para aplicar estas herramientas a sus proyectos de investigación durante la cursada.

Carga horaria total: 64 hs

Carga horaria Clases Teóricas: 32 horas
Carga horaria Clases Prácticas: 32 horas

Metodología de evaluación:

Elaboración de un trabajo final donde se utilicen al menos cuatro de las herramientas aprendidas durante la cursada a la resolución de un problema de análisis de estructura-función de proteínas.

Presentación y defensa del trabajo final.

Programa analítico y contenidos mínimos:

Programa Teórico

1- Bases de datos y anotación funcional de proteínas

Bases de datos para el estudio de proteínas. Exploración de información existente en las bases de datos UNIPROT, INTERPRO y PFAM. Definición de dominios proteicos, sitios funcionales y localización celular. Análisis y obtención de secuencias proteicas.

2- Conceptos de estructura-función de proteínas: Proteínas globulares versus proteínas Intrínsecamente desordenadas (IDPs).

Definición de proteínas intrínsecamente desordenadas (IDPs) y su importancia en las redes de señalización. Graduación de propiedades estructurales e importancia de la flexibilidad en proteínas. Herramientas experimentales para el análisis y estudio de IDPs. Elementos funcionales dentro de regiones IDPs: Motivos Lineales. Análisis e importancia de los motivos lineales en la señalización celular.

3- La modularidad como concepto clave para la comprensión de relaciones estructura-función de proteínas

Conceptos de redes de señalización celulares. Importancia de la localización celular de complejos como determinante de la especificidad de la señalización. Proteínas modulares e interacciones proteína-proteína.

4- Estructura de proteínas. Características y predicción de dominios globulares.

Abundancia de regiones globulares en el proteoma de diferentes organismos. Diversidad estructural en proteínas. Predicción estructural mediante el modelado por homología. Nuevos métodos de predicción estructural: AlphaFold. Conservación de secuencia y estructura.

5- Prácticas experimentales para el estudio de proteínas e interacciones moleculares

Prácticas experimentales para el estudio de interacciones proteína-proteína. Análisis crítico de técnicas *in celula*, sus alcances y limitaciones. Artefactos derivados de la sobre-expresión de proteínas en sistemas eucariotas y procariotas.

6- Detección y predicción de regiones desordenadas en proteínas

Concepto de desorden intrínseco y herramientas experimentales y bioinformáticas utilizadas para la detección y predicción de regiones IDPs. Bases de datos DISPROT y MobiDB, análisis de predictores de desorden utilizando IUPRED y MobiDB. Correlación de la información disponible en las diferentes bases de datos.

7- Detección y predicción de motivos lineales en proteínas

Detección, predicción y análisis de motivos lineales en proteínas. Presentación de la base de datos ELM (Eukaryotic Linear Motif Database). Conceptos teóricos: ¿cómo se define un motivo lineal a partir de evidencia experimental? ¿Cómo se analizan los patrones de conservación de secuencia para identificar motivos lineales? Concepto de expresión regular. Exploración y análisis de proteínas utilizando la base de datos ELM. Algoritmos de predicción de motivos lineales como ANCHOR y bases de datos como ProViz.

Programa Práctico

Trabajo Práctico 1: Bases de Datos de Proteínas
Bases de datos de proteínas. Uniprot. Interpro, PDB, PFAM. y predictores funcionales como TMHMM. Material curado y no curado. Anotación de proteínas y tipos de evidencias.

Trabajo Práctico 2: Visualización de Proteínas en ChimeraX
Entrenamiento en el uso del programa Chimera para la visualización de estructuras de proteínas. Formato de archivo PDB. Comandos para la visualización de proteínas a diferentes niveles.

Trabajo Práctico 3: Modelado por Homología I - HHPred
Uso del programa HHPred para la predicción de estructuras de proteínas a partir de secuencia utilizando templados. Análisis de calidad del modelado y aspectos críticos del modelado por homología. Análisis de modelos utilizando Chimera y comparación con el templado mediante el cálculo de RMSD. Alineamiento estructural.

Trabajo Práctico 4: Modelado por Homología II - Alphafold
Uso del programa Alphafold para la predicción de estructuras de proteínas utilizando ColabFold. Predicción de familias de estructuras. Predicción con y sin templado y comparación/alineamiento estructural. Análisis de conservación utilizando CONSURF y representación de conservación en la estructura de una proteína utilizando Chimera.

Trabajo Práctico 5: Interacciones proteína-proteína
Bases de datos de interacciones proteína-proteína (IntAct, MINT). Análisis de niveles de evidencia en la anotación experimental de interacciones proteína-proteína, evidencia directa e indirecta. Redes de interacción (KEGG y STRING). Visualización de interfaz de interacción proteína-proteína y conservación de sitios de interacción utilizando CONSURF.

Trabajo Práctico 6: Predicción de regiones flexibles (desorden intrínseco)
Uso de diferentes predictores y meta-predictores para el análisis de regiones flexibles de proteínas (IUPRED, MOBIDB). Comparación de predictores de desorden y predictores de estructura (IUPRED versus Alphafold). Análisis de composición de secuencia de regiones desordenadas. Análisis e identificación de regiones desordenadas en alineamientos múltiples de secuencia utilizando Jalview.

Trabajo Práctico 7: Motivos lineales I. Expresiones regulares e identificación en conjuntos de secuencias.
Uso de expresiones regulares (Regex) para la descripción de motivos lineales. Gramática de expresiones regulares y construcción de expresiones regulares que representan a un motivo lineal. Uso de logos de secuencia para el análisis de motivos lineales. Comparación de Regex y logos de secuencia. Predicción de motivos lineales comunes a un conjunto de secuencias.

Trabajo Práctico 8: Motivos lineales II. Análisis estructural de motivos lineales, identificación en alineamientos múltiples de secuencia y mediante la base de datos ELM.
Análisis estructural de motivos lineales y relaciones estructura-secuencia. Análisis de motivos lineales en alineamientos múltiples de secuencia. Predicción y análisis de motivos lineales utilizando la base de datos ELMdb.

Trabajo Práctico 9: TP Integrador
Utilización integrativa de las herramientas aprendidas para el análisis avanzado de problemas de estructura-función e interacciones de proteínas.

Objetivos específicos del curso
Al finalizar el curso, los alumnos/as serán capaces de:

- Comprender aspectos fundamentales sobre el tipo de estructura presente en proteínas eucariotas y procariotas y la importancia de la estructura modular de las proteínas en sus funciones.
- Comprender el tipo de información contenida en las bases de datos de proteínas UNIPROT, PFAM y RCSB-PDB, distinguiendo información curada de información anotada de modo automático.
- Visualizar, representar y analizar proteínas utilizando el programa Chimera (el uso del programa se practicará a lo largo de múltiples TP de la materia)
- Comprender el concepto de modelado por homología y el estado del arte de estas técnicas
- Generar modelos estructurales de proteínas mediante modelado por homología y evaluar la calidad de los modelos obtenidos. Alcances y limitaciones de las técnicas de modelado estructural.
- Realizar comparaciones entre múltiples modelos o estructuras de proteínas mediante alineamiento estructural. Interpretar parámetros de similitud estructural (RMSD) a nivel global y local para definir regiones estructuralmente conservadas.
- Identificar residuos conservados en proteínas mediante la herramienta CONSURF y representar la conservación sobre la estructura conocida o predicha de proteínas.
- Identificar interactores de una proteína de interés utilizando bases de datos de referencia, distinguiendo niveles de evidencia experimental. Análisis de interactores (enriquecimiento funcional), evidencia directa e indirecta.
- Adquirir conceptos de desorden en proteínas y de las diferencias entre regiones desordenadas y regiones globulares. Funciones y ventajas adaptativas de proteínas desordenadas.
- Identificar regiones desordenadas en proteínas de interés a partir del uso de bases de datos (evidencia experimental) y predictores. Evaluación crítica de los resultados. Poder y limitación de las herramientas predictivas.
- Identificar elementos funcionales presentes en regiones desordenadas mediante análisis de conservación y predictores.
- Adquirir conceptos acerca de motivos lineales como una clase funcional presente en regiones desordenadas, y su importancia en las redes de señalización celular.
- Reconocer las técnicas experimentales utilizadas para la validación de interacciones proteína-proteína y motivos lineales.
- Definir motivos funcionales (motivos lineales) a partir de expresiones regulares. Análisis de logos de secuencia. Construcción de expresiones regulares a partir de evidencia experimental.
- Analizar regiones globulares, desordenadas y motivos lineales mediante la inspección de alineamientos múltiples de secuencia utilizando el programa JalView.
- Explorar y predecir motivos lineales en proteínas de interés en la base de datos ELMdb.

Bibliografía:

Libros:

Biochemistry, 5th Ed. Lubert Stryer, Jeremy M. Berg & John L. Tymoczo. W.H. Freeman & Co.

Introduction to Protein Structure, 2nd Ed. Carl Branden & John Tooze. Garland Publishing.

Structure and Function of Intrinsically Disordered Proteins. P. Tompa Ed. Chapman & Hall/CRC Press.

Publicaciones científicas:

- **Estructura y función de proteínas, clasificación de dominios y Modelado por homología:**

Das S, Dawson NL, Orengo CA. Diversity in protein domain superfamilies. Curr. Op. Genet. Dev. 2015 35:40-9. PMID: 26451979.

Todd A.E., Orengo C.A., Thornton, J.M. Evolution of Function in protein superfamilies, from a structural perspective. J. Mol. Biol. 2001 307(4):1113-43. PMID: 11286560

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol.* 2018 Jul 20. S0022-2836(17)30587-9. PMID: 29258817

Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics.* Dec;72(1):e108. doi: 10.1002/cpbi.108. PMID: 33315308

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S and Steinegger M. (2021). ColabFold - Making protein folding accessible to all. bioRxiv. doi: 10.1101/2021.08.15.456425

Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583-589. doi: 10.1038/s41586-021-03819-2. PMID: 34265844

Baek M. et. al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):871-876. PMID: 34282049

- ***Dominios de señalización, proteínas modulares y "scaffolds"***

Scott, J.D. & Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. Science. Nov 27;326(5957):1220-4. PMID: 19965465

Langeberg, L.K., Scott, J.D. (2015). Signalling scaffolds and local organization of cellular behaviour. Nat Rev Mol Cell Biol. Apr;16(4):232-44. PMID: 25785716

- ***Desorden proteico***

Wright PE, Dyson HJ. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 293(2):321-331. doi:10.1006/jmbi.1999.3110. PMID: 10550212

Uversky V.N, Gillespie, J.R., Fink, A.L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41:415–427. PMID: 11025552

Dyson, H.J. & Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. Mar;6(3):197-208. PMID: 15738986

Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z and Uversky VN. The unfoldomics decade: an update on intrinsically disordered proteins. PMID: 18831774

Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci. Dec;37(12):509-16. PMID: 22989858

Oldfield, C.J. & Dunker, A.K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 83:553-84. PMID: 24606139

van der Lee R, Buljan M, Lang B, et al. (2014). Classification of intrinsically disordered regions and proteins. *Chem Rev*. 114(13):6589-6631. doi:10.1021/cr400525m. PMID: 24773235

- **Propiedades de secuencia de regiones flexibles**

Marsh JA, Forman-Kay JD. (2010). Sequence determinants of compaction in intrinsically disordered proteins. *Biophys J*. 98(10):2383-2390. doi:10.1016/j.bpj.2010.02.006. PMID: 20483348.

Das RK, Pappu R V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A*; 110(33):13392-13397. doi:10.1073/pnas.1304749110. PMID: 23901099.

Das RK, Ruff KM, Pappu R V. (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol*. 32:102-112. doi:10.1016/j.sbi.2015.03.008. PMID: 25863585

Mier P. et. al. (2020). Disentangling the complexity of low complexity proteins. Brief. Bioinform. 21(2):458-472. PMID: 30698641

Lupas, A.N., Bassler, J. (2017). Coiled Coils- A model system for the 21st century. Trends Biochem Sci. 42(2):130-140. PMID: 27884598

- ***Motivos Lineales***

Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. Mol Cell. Jul 17;55(2):161-9. PMID: 25038412

Davey, N.E., et al. (2012). Attributes of short linear motifs. Mol Biosyst. Jan;8(1):268-81. PMID: 21909575.

Via, A., et al. (2015) How pathogens use linear motifs to perturb host cell networks. Trends Biochem Sci. Jan;40(1):36-48. PMID: 25475989

Davey, N.E., Travé, G., Gibson, T.J. (2010). How viruses hijack cell regulation. Trends Biochem Sci. Mar;36(3):159-69. PMID: 21146412

- ***Métodos experimentales para el estudio de la función de proteínas***

Gibson, T.J. et al., Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. Cell Commun Signal. Nov 18;13:42. PMID: 26581338

Gibson, T.J., Seiler, M., Veitia, R.A. (2013) The transience of transient overexpression Nat Methods. Aug;10(8):715-21. PMID: 23900254

Greenspan N.S. (2011). Attributing functions to genes and gene products. Trends Biochem Sci. 2011 Jun;36(6):293-7. PMID: 21269834

- **Bases de datos y software cubiertos en la materia:**

- UniProt: http://www.uniprot.org/
- Pfam: http://pfam.xfam.org/
- PDB: https://www.wwpdb.org/
  https://www.rcsb.org/
  https://www.ebi.ac.uk/pdbe/
- InterPro: https://www.ebi.ac.uk/interpro/
- TMHMM: https://services.healthtech.dtu.dk/service.php?TMHMM-2.0
- JPred: http://www.compbio.dundee.ac.uk/jpred/
- IUPred: http://iupred2a.elte.hu/
- Anchor: http://anchor.enzim.hu/
- DisProt: http://www.disprot.org/
- MobiDB: https://mobidb.bio.unipd.it/
- ELM: http://elm.eu.org/
- SLiMSearch: http://slim.ucd.ie/slimsearch/
- ProViz: http://proviz.ucd.ie/
- Jalview: http://www.jalview.org/
- UCSF Chimera: https://www.cgl.ucsf.edu/chimera/
- Reactome: https://reactome.org/
- KEGG: http://www.genome.jp/kegg/
- PED: https://proteinensemble.org/
- CIDER: http://pappulab.wustl.edu/CIDER/analysis/
- HHPred: https://toolkit.tuebingen.mpg.de/tools/hhpred
- Alphafold-ColabFold:
  https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
- Sequence Logos: http://weblogo.berkeley.edu/logo.cgi
- Regex101: https://regex101.com/