Seminarios de Investigación

Resumen

Durante esta asignatura se irán presentando distintos investigadores ponentes para explicar sus investigaciones y aportaciones bioinformáticas.

Sandra Mingo Ramírez

UAM - 2025/26

24 de octubre de 2025 17:35

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

.1. Bioinformática en la investigación dentro de institutos de investigación sanitaria

Val Fernández - investigador Instituto Ramón y Cajal de Investigación Sanitaria

Hay dos grandes fuentes en ciencia para dar financiación: ministerio de ciencia (agencia estatal de investigación) y ministerio de sanidad (Instituto de Salud Carlos III). Este último financia sólo proyectos relacionados con la sanidad, el otro no. La agencia estatal financia todo, denominado como "plan nacional". El ISCIII tiene una cogobernanza con el ministerio de ciencia. Pueden pedir financiación IIS y CNIO/CNIC/CNE/etc y todos aquellos proyectos de universidades y CSIC siempre y cuando estén asociados a un IIS (FIS).

Dentro del hospital están los IIS: IRYCIS, IdiPaz, etc. Pero también hay algunos grupos concretos dentro del IIS que están asociados a universidades. Dentro de los hospitales y los IIS están las fundaciones, que son los encargados de gestionar el dinero de investigación y que no se mezcle con el dinero destinado a tratar pacientes.

Puede haber bioinformáticos clínicos contratados por el hospital y bioinformáticos contratados por la fundación. Esto hace que los bioinformáticos contratados no deberían hacer labor asistencial. En general se divide:

- Bioinformático en IIS: realiza investigación en un área temática determinada en un grupo de investigación.
- Bioinformático clínico: realiza análisis bioinformáticos destinados a la actividad asistencial. Un bioinformático puede trabajar en la genética médica, oncología, hematología, microbiología, enfermedades infecciosas, inmunología, unidad de data science, etc. En la parte asistencial, la realidad es que depende de cada hospital. Normalmente los bioinformáticos clínicos están asociados a la genética médica, y la oncología está algo atrasada en cuanto a la secuenciación masiva.
- Bioinformático servicio: proporciona soporte bioinformático a los investigadores del instituto. Esto se cobra de los proyectos, pero es una unidad central a la que poder pedir soporte y ayuda. Entre todas las labores que se hacen es apoyo a las ómicas, asesoramiento y ayuda en el diseño de proyectos que incluyan análisis bioinformático, data mining, desarrollo de herramientas de análisis, formación del personal del IIS, etc. Se suele dar servicio a todos los grupos que no tengan un bioinformático. Cuando un grupo se va adentrando en un tema, suelen contratar a un bioinformático y no acuden al servicio. En algunos campos, todos los investigadores tendrán que aprender algo de bioinformática para poder al menos analizar sus datos.

Las ómicas siempre son distintas en los grupos de ciencia y los del hospital. No se puede calcular un tamaño muestral, se realizan análisis multivariante en donde se suele buscar qué variable nos sirve, y lo más importante, trabajamos con lo que podemos. Pero existen unos mínimos.

Se debe tener en cuenta la variabilidad técnica-muestra. El problema de la muestra humana es que hay muchas variables distintas entre todos, y se debe tener muy en

cuenta. Por ello, la complejidad del estudio en términos de grupos a comparar debe ser un balance entre ambición y realidad. Los grupos control son esenciales, y en ocasiones se necesita más de un grupo control. Algunas ómicas son dependientes de referencia y otras independientes, y algunas tienen referencias fácilmente buscables en el NCBI o bases de datos similares.

En el análisis, se debe saber si existe un consenso en la comunidad, si existen pipelines estandarizados (en nf-core o similares) y si hay infraestructura para analizarlo.

Un bioinformático de servicio debe saber de todo: análisis de secuencias, anotaciones de genomas, análisis de la expresión génica, análisis de la regulación, análisis de mutaciones, predicción de la estructura de las proteínas, genómica comparativa, modelado de sistemas biológicos, análisis de imagen de alto rendimiento, acoplamiento proteína-proteína, etc.

Mensajes más importantes: "hacemos lo que podemos con lo que tenemos" y "te aviso cuando esté listo":)

.2. Comparative primate genomics to understand evolution, health, and disease

David Juan - investigador Centro Nacional de Biotecnología

Primates are one of the model systems. The closest relatives of humans are chimpanzees and bonobos, although the most commonly used model animal is the mouse. Primates have a higher relevance for disease, immunity, cognition and aging, unique phenotypes absent in rodents like longevity, brain development and cancer resistance, and more similar genomes and transcriptomic and epigenomic regulation. However, working with primates is very difficult due to ethical and legal restrictions, limited availability, high maintenance cost, slower breeding cycles, fewer genetic tools and established inbred lines compared to mice and smaller sample sizes reduce statistical power.

In 2023, we had 20 long-read and 40 short-read assemblies, in addition to some RNA-seq data for almost all these species. Epigenomic studies exist only for great apes and a few other primates. Most studies are focused on the evolution of the human brain, why humans are so intelligent. That being said, omic data per tissue is scarse, mostly concentrated on the brain.

In 2023, 17 new long-read assemblies were published together with 233 primate short-read genomes (assemblies and several individuals).

The group is focusing on great apes to understand human mutational processes. To study aberrant mutational processes in tumors, they compare them to normal germline mutations extracted from genomic population data.

Populations of great apes are better models for understanding the accumulation of somatic mutations in human cells than human population. In fact, gorillas and chimpanzees are more correlated to human tumors than humans themselves.

.3. Gut microbiota-derived metabolites: discovery of biomarkers and therapeutic targets in CVDs

Annalaura Mastrangelo - Immunology Lab CNIC

Atherosclerosis (AT) is a silent precursor of cardiovascular diseases (CVDs). Traditional cardiovascular risk factor-based scores fail to identify individuals at risk at early stages. The rate of CV events remains high despite good cholesterol control. The only treatments available today are lipid-lowering. However, microbiota-host crosstalk has been suggested as a contributor to AT. Microbial imidazole propionate (ImP) is associated with complex diseases and all-causes mortality like Alzheimer's.

How does the gut microbiota affect AT? A murine model was used with different diets and antibiotics after 4 weeks. Antibiotics were able to decrease the progression of atherosclerosis. Plasma metabolome is altered by diet and antibiotics, and the microbiome diversity is reduced in HC diet. The metabolite TMAO was already associated to AT in the literature, but ImP, a microbial metabolite, was also associated in mice.

Higher plasmatic ImP is independently and strongly associated with subclinical AT, particularly active AT. ImP shows additive value when included to established AT biomarkers. To validate the biomarker, the patophysiology of AT and the role of ImP must be studied, as well as if there is a causal role.

Two murine models with AT were used to see if the metabolite alone was able to induce the disease. ImP in drinking water was able to induce the disease in both models. So, ImP induces AT in proAT mice fed chow diet without affecting cholesterol levels in plasma. In the blood, ImP administration expanded proinflammatory Ly6C high monocytes, T-helper 17 (Th17) and Th1 cells. ImP administration induced an increase in fibroblasts, endotelial cells and immune cells, particularly T and B cells.

ImP exerts its role on its targets cell by acting on the imidazoline 1 receptor (I1R Nisch), which is blocked by AGN192403. In vivo, ImP drived atherosclerosis via I1R in myeloid cells. AGN was able to prevent AT progression upon high-cholesterol disease.

ImP is associated with AT in mice and humans, possibly serving as a biomarker of early and active AT. ImP alone induces AT by activating proatherogenic systemic innate and adaptative responses, without influencing bloodstream cholesterol.

The ongoing project is to test synergistic therapy with lipid-lowering treatments and generation of new molecules blocking the ImP/I1R axis. In addition, the development of a MS-based diagnostic tool to establish robust, standardized MS-based methodologies and ready-to-use kits for the reliable and reproducible quantification of ImP in biofluids for clinical applications. In clinics, it is important to define the use of high ImP as a marker for prognosis of future cardiovascular events and define physiological and pathological ranges of ImP in biofluids, together with testing the novel diagnostic devices for ImP quantification in clinical samples.

.4. Introduction to the use of GitHub and its applications in bioinformatics

Adrián Martín Segura - IMDEA Nutrición

GitHub Pages allows to host websites for free.

Git is an open-source version control system, a software to track the changes in code, manage files and directories and revert the changes. Git Bash is an easy way to use git on Windows. GitHub is an online hosting service, the cloud for git.

The advantage of GitHub is to be able to have different versions and merge them into a common one. Basic GitHub terms:

- Clone: making a local copy of a repository
- Commit: register the changes from a file
- Pull: take the main branch to the local copy
- Push: take the recent commits to the main branch of the remote from the local copy

If you clone a repository, you create a local copy. If you are not the owner of the repo, you have to ask permition every time you want to push (pull request). If you do a fork, you create a remote copy to your GitHub.

```
cd route/to/your/directory
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin https://github.com/username/repo_name.git
git push -u origin main

git clone https://github.com/username/repo_name.git

git branch branch_name
git checkout branch_name
# Alternativa en un comando: git checkout -b branch_name

git branch
nano README.md
git add README.md
git commit -m "Update README via terminal"
git push
```

.5. Personalized medicine: An achievable challenge to tackle with AI

Álvaro Serrano - CNIC

Dentro de la medicina personalizada está la farmacogenómica, que es el estudio de cómo los genes de un individuo determinan su respuesta a fármacos y generar tratamientos individualizados (fármaco adecuado en dosis adecuada). Así, se busca conseguir mejores resultados con menos efectos secundarios.

En el CNIC hay un caso de estudio del receptor ADRB1 betaadrenérgico. Se une a catecolaminas e inicia la cascada de señalización. Hay diversos fármacos betabloqueantes que modulan su actividad. Se diferencian por su selectividad, afinidad y si modifican el cambio conformacional que da lugar a la señal. El único que produce un cambio de conformación es el metropolol.

El receptor tiene mutaciones descritas, como Arg389Gly. Los individuos con esta mutación, que no siempre es patogénica, tienen un receptor más activo que va a señalizar mejor cuando se une la catecolamina al producir el cambio conformacional.

A día de hoy se utiliza virtual screening. Antiguamente se tenían placas gigantes y robots que iban poniendo librerías de fármacos en los pocillos para ver si los fármacos tenían función. El virtual screening simula eso con métodos computacionales. La ventaja es que no hay limitaciones físicas y se pueden testar miles de moléculas. Se encuentran compuestos prometedores que sí se prueban en un laboratorio de forma reducida. Así se reducen costes, tiempo y animales de experimentación.

Dentro del virtual screening, atendiendo al conocimiento de la proteína y del ligando, se diferencian las aproximaciones centradas en el ligando, aproximaciones centradas en la estructura y aproximaciones de novo.

En general, los pasos a seguir son:

1. Preparar la molécula diana, obtener la estructura tridimensional con PDB, etc.
2. Preparar el ligando, obtener su estructura tridimensional (en las bases de datos suelen estar bidimensionales o incluso unidimensionales)
3. Docking molecular, ver cómo encajan estas estructuras, evaluar las posiciones y hacer un ranking
4. Elaborar una lista de hits con moléculas en un rango determinado (para poder administrarlo) que son susceptibles de ser un fármaco y estudiar sus propiedades farmacocinéticas y fisicoquímicas.

Teniendo el ligando original, se codifica utilizando descriptores moleculares, una huella que va a codificar las propiedades que tienen. Así se permite buscar de forma eficiente en las librerías de compuestos al permitir filtrar.

En la aproximación basada en estructura, se busca determinar la cavidad en la que van a interaccionar las drogas. Primero se caracterizan y luego se sacan los compuestos para realizar el docking molecular. Se predice la afinidad, energía y demás mediante herramientas como AlphaFold. Características de la proteína como volumen,

polaridad, puentes de hidrógeno, orientación de cadenas laterales y SASA permiten buscar características complementarias de los fármacos para obtener candidatos.

La etapa final del virtual screening es la simulación de dinámica molecular. Al final, cuando se hace una etapa de docking, se obtiene una imagen fija, pero no se sabe si el sistema es termodinámicamente estable. Por ello se ve cómo evoluciona el sistema en el tiempo en base a propiedades fisicoquímicas. Se puede calcular la constante de disociación para ver la concentración a la que se produce la interacción y la cantidad del fármaco que se tiene que suministrar al paciente.

¿Qué pasa si en estos pasos de virtual screening se introducen distintas inteligencias artificiales o modelos de aprendizaje profundo? Desde 2018, estas herramientas han pasado de predecir estructuras con una precisión del 70 % a un 99 % con respecto al PDB. Esto supuso que se le diera el premio Nobel a los creadores de AlphaFold en 2024.

AlphaFold aprende de la evolución, lo que quiere decir que procesa las estructuras del PDB y realiza alineamientos múltiples de secuencia. De ellos, infiere coevolución, ve residuos que van a mantenerse en posiciones concretas a lo largo de la evolución. Esto se representa con una matriz que relaciona la coevolución con la distancia de los aminoácidos, y sobre eso genera unas restricciones geométricas. Con una red neuronal de atención genera las proteínas. El evoformer coge la información del MSA y las restricciones geométricas. Sobre el MSA, utilizando la red neuronal, aprende patrones que le sirven para inferir con otras secuencias las restricciones. El módulo triangular self-attention genera nuevas distancias para generar una estructura 3D. El IPA mantiene las relaciones internas para que sean independientes a las coordinadas absolutas. Por último, se reciclan las predicciones para volver a alimentar la red y crear restricciones más precisas.

No solo nos interesa la estructura de la proteína, si no la interacción entre proteínas o entre proteína y el fármaco. Boltz2 es una herramienta que permite predecir un complejo con perturbación de energía libre. Los métodos de perturbación de energía libre clásicos hacían una simulación en un espacio con propiedades fisicoquímicas que ya se conocían. Boltz2 aprende las reglas de perturbación para ser capaz de predecir la interacción. Este método duplica la precisión de otros métodos basados en ML. La IA nos está permitiendo, desde una secuencia de aminoácidos de una proteína y la secuencia unidimensional de un fármaco, la estructura tridimensional del complejo y sus interacciones.

Conociendo la cavidad, se puede crear el fármaco dentro. Para ello se utilizan distintos tipos de arquitectura: VAE, GAN y modelos de difusión. VAE genera variaciones sobre el input original. De esta forma, saca moléculas que varían entre sí. GAN genera dos redes, una que genera formas aleatorias y otra que selecciona las buenas. Los modelos de difusión son los que mejor funcionan, pero más costosos son computacionalmente. Sobre conjuntos conocidos se aplica ruido que va difuminando y quitando ruido para generar nuevas moléculas. Es el que mejor resultado está dando. Otros modelos son conjuntos, modelando la interacción directamente en lugar de tener un módulo para la proteína y otro para los fármacos.

Esto abre la puerta a la medicina personalizada a través de la farmacogenómica. Hay varios artículos recientes en los que se han descrito proteínas y fármacos mediante IA.

.6. Los límites de computación en bioinformática

Alfonso Valencia - BSC-CNS

.6.1. Computación

Un superordenador es un ordenador con grandes prestaciones. En el caso del BSC, son dos ordenadores: uno de general purpose y otro de uso acelerado. Esto es parte de una red europea de ordenadores grandes.

Ahora mismo hay 12 factorías de inteligencia artificial financiadas en la Unión Europea. Estas factorías pretenden dar soporte al desarrollo de modelos, incluidos los conjuntos de datos, los cuales contienen datos médicos, imagen médica, datos genómicos, trayectorias de dinámica molecular y librerías de pequeñas moléculas.

.6.2. Gemelos digitales en el BSC

El gemelo digital corre a la vez que la factoría y simula todo. Esto sirve para organizar la forma más rápida de reestablecer la cadena de fabricación en caso de que falle un robot. Hay otros gemelos digitales: un gemelo digital de la Tierra para predecir el cambio climático, gemelos digitales de las ciudades, gemelos digitales de humanos de partes concretas. Se pueden hacer gemelos digitales de modelos moleculares, simulaciones de órganos, de tumores.

Se toman datos de single cell para tener la información de señalización celular y sus genes y proteínas activadas. Con esto se pueden simular tumores y organoides. La FDA ha empezado a aprobar medicamentos para humanos en los que parte del dossier se incluyen simulaciones, por lo que ha aumentado el interés en estas técnicas.

El problema de los gemelos digitales en humanos es que es un sistema multiescala, son simulaciones a gran escala y la definición de parámetros y estructura interna en la señalización celular.

.6.3. Acceso a datos

La promesa del futuro es poder acceder a los datos, que estén disponibles para poder inferir el mecanismo celular y patológico. El mayor experimento que puede tener la humanidad es en la práctica clínica con los sujetos que van al hospital y se someten a tratamientos, pruebas, imágenes, etc. Todos los datos que se generan en la atención primaria recientemente se deben poder utilizar para la investigación según una normativa europea. Hay varias iniciativas en Europa para que los datos sean accesibles sin que salgan desde donde están: Elixir, Eucaim, IMPaCT.

Hay un sistema federado probado en distintos sitios en España que permiten acceder a unas herramientas y potencia computacional con unos datos sin que salgan del sitio que se conecta. El problema a nivel europeo es que la burocracia limita todo y no fue posible acceder a los datos debido a que las autoridades de los datos no lo permitieron. No obstante, el siguiente problema será la (falta de) interoperabilidad de los datos de

los distintos sitios. A futuro se considera la ontología de OMOP o una curación del texto escrito por médicos a la terminología de la ontología.

.6.4. Inteligencia artificial

Las aplicaciones más interesantes que están surgiendo en la inteligencia artificial son en el campo de la biomedicina al ser el campo más complejo. Hay modelos de lenguaje específicos para la medicina en la que se extraen los conceptos del texto del historial médico. Con esto se busca sacar unos guidelines para finales de año.

El mayor logro de la IA fue el modelado molecular de proteínas. Esto permite no solo crear nuevos fármacos a unas velocidades desorbitantes, si no también explorar la evolución de las funciones proteicas y explorar el espacio de secuencias. Además, esto permite generar datos sintéticos que se pueden utilizar para modelar. No obstante, ahora entra la pregunta, ¿cómo sabemos que los datos sintéticos son correctos? Con las proteínas era fácil saberlo por cristalografía de experimentos previos.

El sesgo más importante que se encuentra es relativo al sexo y género. La gran mayoría de los datos se han obtenido de hombres o ratones machos, y en el caso de los humanos suele ser hombres blancos de un cierto rango de edad. Esto es terrible porque se ha estudiado que hombres y mujeres tienen pathways distintos para el dolor y las patologías. No obstante, también hay otros sesgos como los técnicos.

La mayor parte de las interacciones por la web son por agentes. Los agentes de IA son sistemas de software autónomos que utilizan IA para conseguir objetivos específicos y realizar tareas por los usuarios. Pueden interaccionar con el entorno y tomar acciones.