BIOINFORMÁTICA, ¿SOLO PROGRAMAS COMPUTACIONALES PARA ALMACENAR DATOS BIOLÓGICOS?

MARIOINOSTROZA

Profesor Asociado Departamento de Ingeniería Informática, Universidad de Santiago de Chile (DIINF-USACH). PhD Computer Science, Universidad de Newcastle, Australia. Líneas de Investigación: Aplicación y desarrollo de técnicas Metaheurísticas, de Optimización Combinatorial, Optimización Multiobjetivo aplicados a problemas en Bioinformática, tales como Análisis de datos de Expresión Génica, Predicción de Estructura de Proteínas, Docking Molecular e Inferencia Filogenética.
mario.inostroza@usach.cl

Durante la educación media, la mayoría de los y las jóvenes deben tomar la decisión sobre qué área seguir. Dependiendo del establecimiento educacional, éstas reciben diferentes denominaciones: matemático, biólogo, humanista, arte y una variedad de otros nombres, que intentan describir cuál es el área de interés que seguirá profundizando durante su educación secundaria. En nuestros días esta decisión puede ser considerada arcaica, toda vez que la educación y las especializaciones tienden a trabajos multidisciplinarios, que hace poco tiempo no existían. El desarrollo de muchas áreas en general se ha dado de manera independiente, pero tienden a juntarse dada la sinergia que se produce al unir una o más disciplinas. Ejemplos claros de esta situación son bioquímica, biofísica, econofísica, computación afectiva, entre otras que aún no han recibido un nombre propio. En particular, con el nacimiento de los primeros computadores, los avances obtenidos en el área de ciencia de la computación y especialmente los rápidos avances tecnológicos en el área de la biología, hemos visto el nacimiento de una nueva área de desarrollo, la bioinformática. Ésta nació como respuesta natural a las nuevas necesidades que la biología presenta por la gran cantidad de datos generados. Sin embargo, lo que en un principio apuntaba solo al manejo de grandes volúmenes de datos, rápidamente se convirtió en un área crucial para el entendimiento de sistemas biológicos complejos, gracias a las técnicas y algoritmos avanzados que han permitido extraer información que a simple vista parecía oculta. En este artículo revisaremos brevemente una perspectiva histórica de la bioinformática, el aporte de la ciencia de la computación a la biología, a través de un ejemplo concreto, y una pequeña discusión sobre cómo es el trabajo entre ambas disciplinas y los diversos roles profesionales que existen en el área.

UN POCO DE HISTORIA

Los primeros antecedentes del uso de la palabra bioinformática se remontan a 1970, en un trabajo de Hesper y Hogeweg donde se define como "el estudio de los procesos de información en sistemas bióticos". Esta definición hace que sea equivalente a conceptos como biofísica (el estudio de procesos físicos en sistemas biológicos) y bioquímica (el estudio de procesos químicos en sistemas biológicos). Sin embargo, desde esa fecha hasta ahora, el concepto de bioinformática ha mudado, debido en parte a su característica multidisciplinar y a los avances tecnológicos que han creado nuevos desafíos. Actualmente, bioinformática puede ser entendida como el área que se preocupa por la recolección y el análisis de datos biológicos complejos, por medio del desarrollo de métodos computacionales y herramientas de software avanzadas. El objetivo final, es el descubrimiento de la información que se encuentra en los datos biológicos recolectados, para apoyar el entendimiento y estudio de los procesos biológicos de interés.

Sin embargo, llegar al estado actual de desarrollo de la bioinformática responde a una serie de avances tanto en ciencias de la computación como en biología. Antes de la década del setenta, hubo grandes avances en biología que permitieron conocer un poco más la forma en que nuestro código genético está escrito y como funciona: la estructura de hélice del ADN (1953), propiedades moleculares estructurales (1953, 1957), rutas metabólicas (1945), regulación génica (1969), entre muchas otras. Por el lado de la ciencia de la computación podemos contar bases teóricas en el ámbito de Teoría de la Información (1962, 1966), Definición de Gramáticas (1959), Teoría de Juegos (1953), Autómatas Celulares (1966) entre otros avances. Luego, en la década del setenta, donde se reconoce el primer uso del término bioinformática, se considera como la década donde se desarrollaron las bases teóricas del área: métodos de análisis de secuencias, predicción de la estructura del RNA, aplicaciones de máxima verosimilitud a inferencia filogenética, algoritmos de predicción de estructura secundaria de proteínas, entre otros. Además en esta década, se crean los primeros repositorios públicos de secuencias de proteínas, lo que puede ser considerado como una característica clave en esta área.

Ya en las décadas del ochenta y noventa, se produce un rápido avance de la disciplina, principalmente por la generación masiva de datos, lo que creó la necesidad de contar con algoritmos eficientes para su procesamiento. En esta época nacieron los primeros softwares comerciales, bases de datos moleculares (GenBank, EMBL Data Library). Se crean algoritmos que atacan los cada día más comunes problemas de comparación de secuencias (algoritmo Smith-Waterman para alineamiento local), y en el área de proteínas nacen los primeros softwares de visualización de estructuras proteicas, de comparación de estructuras y de estrategias de predicción de estructura terciaria de proteínas. En particular este último continúa siendo hasta el día de hoy uno de los problemas más difíciles de resolver y para el que se han dedicado muchas de las máquinas de la lista del top500 (https://www.top500.org/). En particular en la década del noventa, el área se vio revolucionada por la creación del algoritmo BLAST (Basic Local Alignment Search Tool), que catalizó los procedimientos de comparación de secuencias y permitió realizar la búsqueda de secuencias objetivo sobre alguna base de datos de referencia. Si bien BLAST, a diferencia del algoritmo Smith-Waterman, no garantiza el mejor resultado de alineamiento, la velocidad de su respuesta al buscar sobre grandes bases de datos de información genómica, es característica suficiente para ser una de las herramientas más usadas en bioinformática.

Por otro lado, el rápido crecimiento de Internet, la masificación de sistemas de cómputo de alto rendimiento y por sobre todo el aumento explosivo de datos biológicos almacenados, provocó la aparición explosiva de nuevas técnicas computacionales que hacían uso de estos sistemas de cómputo y procesaban los datos libremente disponibles a través de las plataformas virtuales. El proyecto GOLD[1] del Joint Institute of Genomics mantiene un repositorio de los proyectos de genoma existentes, donde puede ser consultado el número de proyectos de secuenciamiento de genomas por año y su referencia, dando una idea clara del crecimiento masivo de los datos disponibles y que necesitan ser analizados. Finalmente, el rápido crecimiento del poder computacional en los últimos quince años, sumado al cambio de paradigma de cómputo y a una disminución dramática en el costo de la tecnología, crean un nuevo escenario, donde los sistemas de análisis de datos ya no solo deben lidiar con datos ruidosos y masivos, sino con nuevos tipos de datos que entregan otra mirada de los procesos biológicos en estudio. La tendencia es hacia el entendimiento de los procesos biológicos a un nivel tal que permita diferenciar entre individuos y poder entregar en un futuro no muy lejano el concepto de medicina personalizada.

¿CÓMO LA CIENCIA DE LA COMPUTACIÓN Y LA INFORMÁTICA AYUDAN A LOS BIÓLOGOS?

Los principales efectos del nacimiento y desarrollo de la bioinformática pueden verse principalmente en los grandes avances en medicina y biología que se han producido, dado que ha aportado al entendimiento de los complejos sistemas biológicos que rigen nuestro diario vivir. Sin embargo, no solo la biología se ha visto beneficiada, sino que la computación y la informática han debido crear nuevas formas de trabajar con estos problemas. Dentro de las técnicas computacionales usadas se pueden encontrar: modelamiento con grafos y sus algoritmos, string matching, uso de gramáticas, optimización lineal, no lineal, optimización combinatorial, estructuras de árbol y sus algoritmos, procesamiento de lenguaje natural, búsqueda de patrones, metaheurísticas, heurísticas, redes neuronales, computación de alto rendimiento, ingeniería de software, diseño de interfaces de usuario (muy importante a la hora de construir herramientas útiles para los biólogos), solo por mencionar algunas. Sin embargo, el uso de estas técnicas y algoritmos ha debido adaptarse a las particularidades del área, donde lo común son las excepciones, más que las reglas. Algo difícil y frustrante para personas que estamos acostumbradas a encontrar un flujo en el procesamiento, en las reglas que gobiernan los procesos y codificarlas de manera elegante en nuestros códigos.

Los datos de expresión génica corresponden a la medición de la actividad de una secuencia génica bajo ciertas condiciones, como enfermedad, tratamiento con medicamentos, una serie de tiempo, etc. Son usados típicamente para

1. Genomes Online Database, https://gold.jgi.doe.gov/

conocer el comportamiento de la generación de productos génicos (como proteínas) sobre un conjunto de muestras biológicas. Para hacer esta medición se usan los microarreglos. Esta tecnología, considerada un tanto antigua al día de hoy, ha permitido conocer con mayor detalle cómo se comportan los genes en un gran número de procesos biológicos, descubrir y anotar genes de acuerdo a sus patrones de expresión, e identificar posibles tratamientos a ciertas enfermedades. Este tipo de datos posee características de ser ruidosos, poseer pocas muestras y contar una gran cantidad de genes a estudiar. Es común que la relación entre el número de genes y el número de muestras sea de varios órdenes de magnitud (por ejemplo, un conjunto de datos podría tener la expresión de 25.000 secuencias sobre 100 muestras). Existen variadas herramientas de análisis de este tipo de datos, siendo la principal el *clustering* o agrupamiento de acuerdo a sus perfiles de expresión. Dentro de los algoritmos de *clustering* más comunes y usados por la comunidad bioinformática, están *k-means*, agrupamiento aglomerativo y *self organizing maps*. Si bien sus resultados en general son satisfactorios, en la literatura es posible encontrar muchos ejemplos de técnicas más avanzadas, no necesariamente más complejas, que producen resultados más completos. Para ejemplificar el uso de una mezcla de estas técnicas, pasaré a explicar un *pipeline* de análisis de datos de expresión génica, basada en grafos, árboles, optimización combinatoria y metaheurísticas.

El pipeline está compuesto de dos partes: (1) el algoritmo de *clustering* MSTkNN y (2) la aplicación de algoritmo QAPgrid para producir un layout de los genes o secuencias en un espacio cartesiano, que responda a las relaciones entre los genes. En la Figura 1 se puede ver gráficamente cómo es el proceso de MSTkNN. Inicialmente se tiene un conjunto de datos de expresión génica y se crea un grafo completo G(V,E,W), donde cada nodo del grafo representa un gen o secuencia, y existe una arista entre todo par de nodos con peso igual a la distancia entre los perfiles de expresión de la secuencia (Proceso (A) en la Figura 1). En este punto es importante destacar que una de las claves en el éxito de aplicación de estos algoritmos está en la definición de la medida de distancia, porque de ella depende lo que el algoritmo considerará como cercano o parecido. Una vez obtenido el grafo completo, el proceso calcula dos grafos de proximidad: un árbol de cobertura mínimo $G(V,E_{MST})$ y un grafo de k vecinos más cercanos $G(V,E_{kNN})$ (proceso (B) en la Figura 1). Posteriormente el algoritmo realiza una intersección entre los conjunto de aristas de esos dos grafos, produciendo una partición del grafo en 1 o más componentes. En caso que se encuentren más componentes, el algoritmo aplica recursivamente el método en cada componente hasta que el grafo no puede ser subdivido. Con este proceso, el algoritmo MSTkNN produce una partición del conjunto de datos sin necesidad de indicar el número de grupos que debe encontrar. Esta característica es importante en esta área, dado que es común desconocer el número de grupos o cómo son las características de los grupos a encontrar. Una vez obtenidos los grupos, es posible generar ahora un layout de los elementos que represente

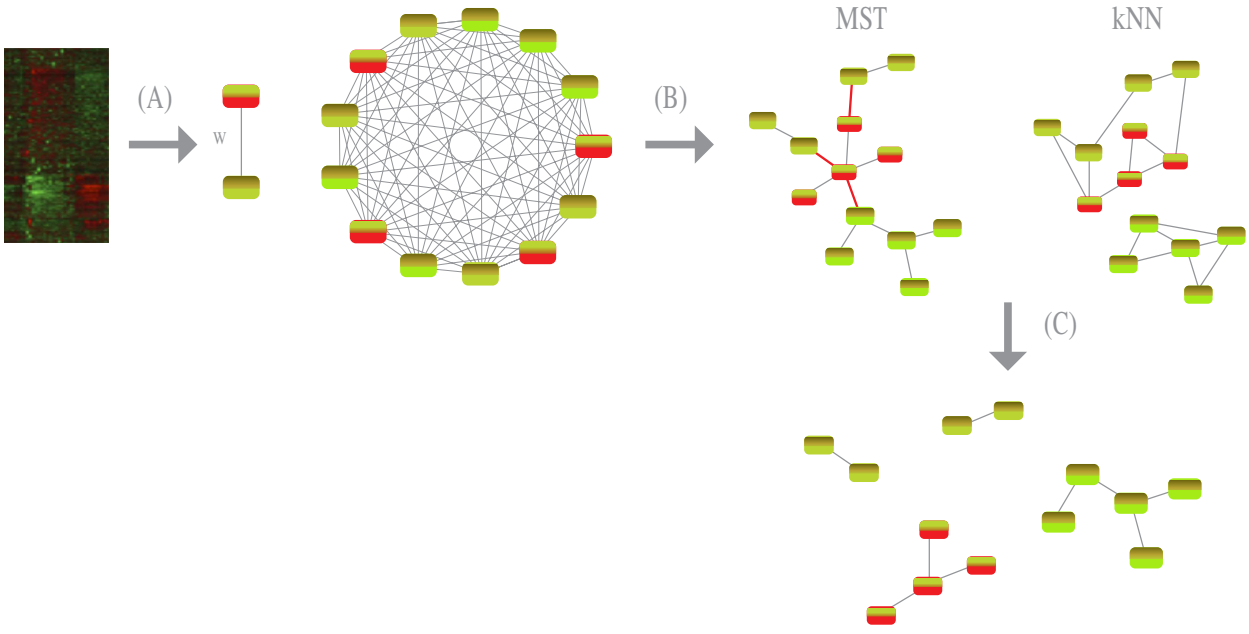

FIGURA 1.
ALGORITMO DE *CLUSTERING* MSTKNN. EL PROCESO SE INICIA CONSTRUYENDO UN GRAFO COMPLETO CON PESOS DE ACUERDO A LA CORRELACIÓN DE LAS EXPRESIONES GÉNICAS. LUEGO INTERSECTA LOS GRAFOS MST Y KNN PRODUCIENDO UNA PARTICIÓN DEL GRAFO ORIGINAL EN UNA O MÁS COMPONENTES.

las relaciones entre ellos. Para esto el algoritmo QAPgrid modela los datos de expresión génica como instancias del problema de asignación cuadrática (QAP por sus siglas en inglés) y usa un algoritmo memético para resolver estas instancias. El problema de QAP es un problema difícil de resolver (pertenece a la clase NP-Hard), por lo que es necesario usar técnicas alternativas, que no garantizan el óptimo pero encuentran soluciones de buena calidad. En la Figura 2 se muestra un esquema simplificado del proceso. Inicialmente se recibe la información de expresión y el resultado de un algoritmo de clustering. Para cada grupo se genera una instancia del problema de QAP (proceso (A) en la Figura 2), más una instancia donde cada objeto del problema de QAP corresponderá a un grupo de la solución del algoritmo de *clustering*. Posteriormente, estas instancias son resueltas por el algoritmo memético mencionado (proceso (B) en la Figura 2). Como resultado se obtendrá un layout de cada grupo en un espacio bidimensional, donde las posiciones de los elementos responden a la similitud entre los perfiles de expresión de cada elemento. Además, cada grupo es también localizado en un espacio bidimensional de acuerdo a la expresión de los elementos miembros de cada grupo, es decir, grupos con elementos parecidos quedarán en regiones próximas del layout. Este método ha sido usado para analizar datos de alzheimer, secuencias no codificantes del cerebro de ratón, datos de expresión génica de muestras de múltiple esclerosis, y otros datos de naturaleza distinta.

La disponibilidad de herramientas computacionales para esta área es gigante, lo que hace fácil caer en la tentación de usar herramientas de manera indiscriminada, sin realmente saber qué y cómo la herramienta enfrenta el problema. Errores típicos de esta área corresponden a la mala parametrización de los métodos y la inadecuada elección de funciones de distancia. Por ejemplo en este último caso, no entender cuál es el impacto de las opciones de distancia empleadas. Malos parámetros igual llevarán a una solución al problema, pero una solución de mala calidad.

UNA MINI TORRE DE BABEL

El trabajo multidisciplinario entre biólogos y científicos de la computación es un desafío en sí mismo. Para poder llegar a las soluciones que existen actualmente se requirió de horas de entendimiento entre personas que han sido formadas de manera distinta, con intereses muchas veces divergentes. Por un lado, la abstracción de los problemas y métodos es una técnica muy usada en computación, que nos permite crear soluciones que abarcan un amplio espectro de casos. Por otro lado, los biólogos están acostumbrados a estudiar procesos particulares y específicos, donde no es solo difícil sino a veces imposible generalizar el funcionamiento o explicación del proceso. Es esta dificultad la que hace fracasar las primeras iniciativas de colaboración. Es común que los científicos más cercanos a la computación no se interesen mucho por

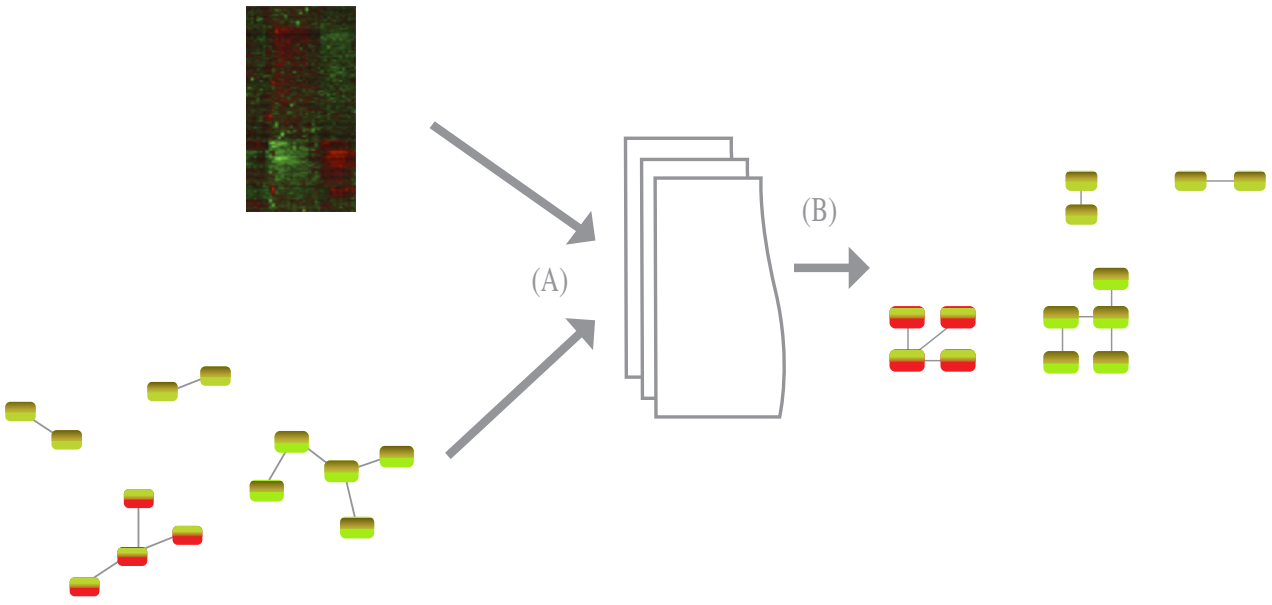

FIGURA 2.
ESQUEMA SIMPLE DE APLICACIÓN DE QAPGRID. RECIBE COMO ENTRADA LAS MATRICES DE EXPRESIÓN GÉNICA Y ALGÚN RESULTADO DE *CLUSTERING* (POR EJEMPLO, EL MOSTRADO EN LA FIGURA 1). LUEGO RESUELVE INSTANCIAS DEL PROBLEMA DE QAP USANDO ALGORITMOS AD HOC PARA ESTE PROBLEMA. LOS OBJETOS SON FINALMENTE ORGANIZADOS DE ACUERDO A SU SIMILITUD DE EXPRESIÓN.

conocer en detalle el problema a resolver y solo se preocupen de los datos. Por otro lado, también es común que los científicos más relacionados al área biológica no se interesen por entender los métodos, ni estén abiertos a nuevas formas de ver o interpretar los datos. Solo quieren "apretar un botón" para procesar sus datos recién recopilados.

Una formación más transversal, con foco en alguna de las áreas específicas (biología o ciencias de la computación) ayudará a formar profesionales que tengan un tiempo de adaptación menor a los diferentes proyectos en los que participen. Para aquellos programas de formación en bioinformática, la participación en proyectos de investigación reales es de gran importancia, así como contar con académicos biólogos que usen herramientas de bioinformática y académicos del área de ciencias de la computación que hayan participado en proyectos de bioinformática. La disponibilidad de recursos online para comenzar en esta área es grande, por lo que es importante tener en cuenta la experiencia de otras instituciones de educación en esta área, y aprovechar los recursos que ofrecen. En este tema, los trabajos de David Searls de 2012 y 2014, dan una luz sobre la oferta académica disponible online a la fecha de publicación. En particular un tópico interesante que se discute en esos artículos tiene que ver con los diferentes roles que una persona en el área puede desarrollar, identificando cinco caminos que tienen que ver con diferentes tipos de desempeño. Esta guía puede ser útil para definir nuestras necesidades como país en la formación de los profesionales que la industria requiere.

COMENTARIOS FINALES

HABLAR SOBRE BIOINFORMÁTICA EN GENERAL NO ES FÁCIL, DADA LA DIVERSIDAD DE ENFOQUES QUE EXISTEN EN EL ÁREA, BASADOS EN LA EXPERIENCIA PERSONAL MÁS QUE EN FUNDAMENTOS BIEN ESTABLECIDOS. OBVIAMENTE UN EJEMPLO DE ESTO ES ESTE ARTÍCULO. EL IMPACTO DE DESARROLLAR FUERTEMENTE ESTA ÁREA PUEDE APOYAR LA CADA VEZ MAYOR NECESIDAD DE DIVERSIFICAR LA INDUSTRIA NACIONAL, CREANDO UN POLO DE DESARROLLO EN ÁREAS COMO LA BIOTECNOLOGÍA, PARA LA QUE SE REQUIERE PROFESIONALES QUE SEAN CAPACES DE ENTENDER ESTE TRABAJO MULTIDISCIPLINARIO. ACTUALMENTE EN CHILE EXISTEN VARIOS GRUPOS DE TRABAJO, CENTROS DE INVESTIGACIÓN DE ALCANCE REGIONAL Y NACIONAL, Y CIENTÍFICOS QUE DESDE LA BIOLOGÍA O DESDE LA CIENCIA DE LA COMPUTACIÓN APORTAN AL DESARROLLO DE LA BIOINFORMÁTICA. ME RESERVO EL DERECHO DE NO NOMBRARLOS PARA NO DEJAR FUERA A NINGUNO DE LOS QUE EXISTEN EN EL PAÍS. FINALMENTE, LA APARICIÓN DE NUEVAS ÁREAS RELACIONADAS A CIENCIAS DE LA COMPUTACIÓN (POR EJEMPLO BIOINFORMÁTICA QUE HEMOS HABLADO EN ESTE ARTÍCULO), VA DE LA MANO CON LAS NUEVAS TECNOLOGÍAS APLICADAS QUE NOS PRESENTAN DESAFÍOS A LOS CUALES DEBEMOS RESPONDER CON NUESTRAS MEJORES PRÁCTICAS TÉCNICAS Y CON LA CERTEZA QUE SE CONSEGUIRÁN MAYORES CONTRIBUCIONES SI SOMOS CAPACES DE ENTENDER EN PROFUNDIDAD LOS PROBLEMAS QUE SE NOS PRESENTAN. ■

REFERENCIAS

Hogeweg P. (2011). The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7(3): e1002021. doi:10.1371/journal.pcbi.1002021.

Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David (1990). "Basic local alignment search tool". Journal of Molecular Biology. 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2. PMID 2231712.

Inostroza-Ponta M, Berretta R, Moscato P (2011) QAPgrid: A Two Level QAP-Based Approach for Large-Scale Data Analysis and Visualization. PLOS ONE 6(1): e14468. doi: 10.1371/journal.pone.0014468.

David B. Searls. An Online Bioinformatics Curriculum PLoS Comput Biol. 2012 Sep; 8(9): e1002632.

Searls DB (2014). A New Online Computational Biology Curriculum. PLoS Comput Biol 10(6): e1003662. doi:10.1371/journal.pcbi.1003662.