- Herramientas Biotecnológicas Como Contribución al Desarrollo de la Biología Evolutiva -

Bioinformática y evolución molecular

Julio Rozas Liras[1,2]

[1]Departament de Genètica. Universitat de Barcelona. Diagonal 645. 08028 Barcelona.
[2]Institut de Recerca de la Biodiversitat de la Universitat de Barcelona.

Introducción

Aunque la bioinformática y la evolución molecular pueden parecer disciplinas científicas muy heterogéneas, comparten gran parte de sus intereses y fundamentos. La bioinformática, por un lado, es una rama de la ciencia que estudia el desarrollo y aplicación de métodos computacionales para resolver problemas de la gestión y análisis de datos biológicos. Mientras que la evolución molecular estudia la dinámica y el proceso del cambio evolutivo a nivel molecular. Estas disciplinas, sin embargo, comparten un mismo interés por el estudio y análisis de secuencias de DNA y proteínas; no sólo eso: estos estudios son el núcleo central de ambas disciplinas. De hecho, varios conceptos fundamentales en bioinformática (alineamiento de secuencias; búsqueda de homólogos; reconstrucción filogenética, etc.) están tan íntimamente relacionados con la evolución que no se pueden entender sin comprender el proceso evolutivo. Y viceversa, no se podría extraer información evolutiva relevante de la ingente cantidad de datos disponible (secuencias de DNA y proteínas de genomas completos) sin el uso de potentes algoritmos y métodos computacionales. Pero es que, además, la bioinformática y la evolución molecular también comparten un origen y una historia común.

Décadas de los 60 y 70

El origen de la evolución molecular se sitúa en los años 60, como una disciplina que surge en la interfaz entre biología molecular, bioquímica, genética, estadística y biología evolutiva. En 1962, Linus Pauling y Emile Zuckerkandl, tras comparar las secuencias de la alfa y beta globina, observan que el número de cambios de amino ácido entre secuencias de diferentes especies es aproximadamente constante con el tiempo de divergencia de las mismas. Por lo tanto, las moléculas de proteínas acumulan información (información evolutiva) que se podría utilizar para medir el tiempo. Esta observación está en la base de uno de los conceptos fundamentales de la

evolución molecular, el reloj molecular. Unos pocos años después Walter M. Fitch y Emanuel Margoliash usaron la información acumulada en la secuencia del citocromo c de varias especies con objeto de inferir la historia de la vida. Con este fin, desarrollan un método de estimación de distancias genéticas entre parejas de secuencias y, además, escriben el primer programa de ordenador para la reconstrucción filogenética que, por ende, representa el primer vínculo entre informática y evolución. Al final de la década de los 60, Motoo Kimura realizó una aportación esencial en evolución molecular: la formulación de la teoría neutralista de la evolución molecular, teoría que entre otras cuestiones proporciona una explicación al concepto de reloj molecular. Según esta teoría la mayoría de mutaciones que ocurren en las poblaciones serían selectivamente neutras (es decir, no proporcionarían ni perjuicio ni beneficio al organismo que las contenga), y cuya frecuencia fluctuaría en las mismas, y eventualmente se podrían fijar, por deriva genética, es decir de forma aleatoria.

Paralelamente Margaret Dayhoff empezó a compilar las secuencias de aminoácidos de las proteínas conocidas. Esta información se publica en el *Atlas of Protein Sequences*, donde se agrupaban las secuencias en familias de proteínas relacionadas, y que se convierte en un claro precursor de las modernas bases de datos. De hecho Dayhoff fue pionera en la utilización de ordenadores para solventar problemas en biología y química, y es por eso que se la considera como la fundadora de la bioinformática. El uso de ordenadores permitió analizar de forma exhaustiva estas secuencias. En particular, el alineamiento de la secuencia aminoacídica entre varias proteínas homólogas sirvió para cuantificar las diferencias que existían entre las mismas, y para estimar las denominadas matrices de substitución PAM (*Percent Accepted Mutations*), en donde se expresa la frecuencia de substitución de un aminoácido por otro. Estos análisis revelaron que, i) no todos los aminoácidos estaban igualmente conservados y, ii) cada familia de proteínas exhibía un nivel y patrón de cambio (evolutivo) característico. Un aspecto sumamente importante para determinar el patrón de cambio aminoacídico radica en la necesidad de identificar correctamente los aminoácidos homólogos; es decir se necesita estimar el alineamiento óptimo entre las secuencias de proteínas. A pesar de que las matrices PAM proporcionaban una metodología para estimar alineamientos, requerían un tiempo de computación formidable. En este sentido fue determinante el desarrollo del algoritmo de Needleman y Wunsch (1970), basado en la técnica de programación dinámica para reducir el tiempo de computación, al posibilitar la realización de alineamientos globales entre dos secuencias en un tiempo razonable.

Décadas de los 80 y 90

En estas décadas se asiste a una tremenda acumulación de secuencias de proteínas, pero sobre todo de DNA. Secuencias de genes y de pequeños genomas completos (virus y procariotas), tanto a nivel intraespecífico como interespecífico. Secuencias que acumulan información sobre los mecanismos evolutivos y la historia filogenética de las especies. Esta información, sin embargo, no habría generado conocimiento científico sin el recíproco avance teórico y tecnológico. Por un lado se asiste al desarrollo de varias metodologías genético-estadísticas para el análisis filogenético; estas metodologías van a permitir utilizar de forma eficiente la información de las secuencias para reconstruir la historia evolutiva de los organismos y estimar los tiempos de divergencia de los principales eventos. Por otro lado, se elabora la teoría de la coalescencia; teoría fundamentada en la genética de poblaciones que facilitará la explotación exhaustiva de la información acumulada en las secuencias. Esta teoría, en particular, proporciona métodos para inferir el impacto de diversos procesos evolutivos (como la selección natural, o eventos demográficos).

Sin el desarrollo paralelo de herramientas bioinformáticas, no obstante, la información presente en las secuencias tampoco habría sido descifrada. Aquí debemos destacar el desarrollo de varios algoritmos computacionales básicos. Por ejemplo, el algoritmo de Smith-Waterman, que permitió obtener alineamientos locales; y cómo no, el algoritmo BLAST (Altschul y col.) muchísimo más rápido y esencial para realizar búsquedas de secuencias por similitud con las existentes en las bases de datos. Tampoco podemos olvidar la gran contribución de la estadística, tanto en la estimación de parámetros, contrastes de hipótesis o simulaciones por ordenador (estimación de parámetros usando modelos ocultos de Markov –HMM- o por máxima verosimilitud; generación de simulaciones por MCMC –*Markov Chain Monte Carlo*-; aplicación de técnicas de inferencia bayesiana, etc).

Años 2000

Se produce la gran eclosión de la genómica. Las nuevas tecnologías experimentales permiten la secuenciación genómica a gran escala, de genomas completos y a un precio razonable. Actualmente ya se dispone de la secuencia completa de unos 2000 virus, 1000 bacterias, 70 arqueobacterias y de unos 30 eucariotas superiores. Estos proyectos no sólo generan datos de la secuencia primaria de DNA, sino también de sus anotaciones (la estructura del gen, su posición en el cromosoma, su posible función, etc.). Y se produce también la eclosión de la genómica comparada en eucariotas superiores. En 2007 se publica la primera comparación y análisis de varios genomas completos de eucariotas superiores (12 especies de *Drosophila*). Y ya están en marcha

proyectos para la secuenciación y análisis de unos 200 genomas de *Drosophila*, y unos 1000 tanto en *Arabidopsis* como en el hombre. Y esto sólo es la punta del iceberg. Indiscutiblemente, la secuenciación en sí misma no es lo primordial, sino el conocimiento científico que puede aportar. Y no nos está defraudando. En pocos años hemos aprendido mucho sobre la estructura y funcionamiento (y mal funcionamiento) de los genes; y de su evolución, sobre los procesos de nacimiento, diferenciación y muerte de los genes (cómo surgen, cómo se diversifican estructural y funcionalmente, cómo se pierden), y sobre los mecanismos evolutivos subyacentes (selección natural positiva y negativa, deriva genética, etc.).

No quiero acabar sin resaltar que este importante conocimiento científico habría sido del todo impensable sin las sinergias generadas por la contribución de varias disciplinas. Se ha necesitado desarrollar aspectos teóricos, analíticos y tecnológicos. Y bioinformáticos, sin cuyas herramientas serían imposibles los denominados proyectos genoma. Y recíprocamente, los múltiples problemas generados por estos proyectos (inimaginables hace unos años) han incentivado la innovación y el desarrollo de nuevas metodologías y tecnologías para solucionarlos. Y sólo a la luz de este abordaje multidisciplinar se puede entender la generación de los nuevos conceptos genético-evolutivos, muchos de ellos inconcebibles hace tan sólo unas décadas. Para finalizar, y en conmemoración de la doble efeméride del año Darwin, me gustaría resaltar que Darwin (1859) no sólo habló de evolución y de selección natural, sino que incluso llegó a hipotetizar la existencia de las mutaciones neutras, uno de los conceptos fundamentales en evolución molecular.

Bibliografía

- C. Darwin. 1859. *On the origin of species*; p 81 (primera edición).

Julio Rozas Liras se doctoró en 1990 en la Universidad de Barcelona. Posteriormente, ejerció como Profesor Ayudante en esta misma Universidad, y tras una estancia Postdoctoral en la Universidad de Harvard, fue Profesor Titular de la Universidad de Barcelona (1992-2009). Actualmente es Catedrático de Genética de dicha Universidad, así como miembro de la Junta directiva del Institut de recerca de la Biodiversitat de la Universidad de Barcelona, Coordinador nacional de REDES (Red Española de Diversidad Biológica, Evolución y Sistemática Molecular), y Miembro del Comité Nacional de IUBS (International Union of Biological Sciences). Asimismo, ha sido Vicedirector del CERTFEM (Centre Especial de Recerca en Taxònomia i Ecologia Moleculars de la Universidad de Barcelona), y secretario de la Sociedad Española de Genética. Es autor de más de 50 publicaciones científicas en temas relacionados con la Genética de Poblaciones, Genómica, Evolución Molecular y Bioinformática.