Odalis J. Jones

Traducción comentada del ámbito biomédico: «La integración de los genomas»

TRABALL FI DE GRAU

JUNIO, 2012

Facultat de Traducció i Interpretació
Universitat Pompeu Fabra
Carles Tebé
Seminari Traducció Cientificotècnica

ABSTRACT

The discipline of Genetics plays a crucial role in the diagnosis, prevention and treatment of many diseases and disorders. The continuous scientific research, especially in this field, makes scientific subjects a frequent and popular source of translation. In this paper we analyze some specialized translation characteristics which become translation problems such as language structure, terminology, and lexical elements. We present these problems by sections together with the translation strategies, the justifications, and the solutions based on the context, field, and target language style in order to better reflect the necessity of linking language skills and scientific knowledge together to produce an accurate and univocal text. However, this work focuses on the translation into Spanish of English complex noun phrases which constitute one of the most troublesome problems to solve within the specialized translation, given the difficulty to analyze the order and connection among noun pre-modifiers. The findings of their translation show obvious signs of the difference not only between languages but also in the way they both describe phenomena and concepts. After a deep reflection on the topics we present in this work, we have been able to provide a professional translation that can be a useful tool to consult and clarify linguistic doubts and to increase the knowledge of genetics and others related disciplines.

TABLA DE CONTENIDO

1. INTRODUCCIÓN _____ *5*

1.1 Objetivos _____ **7**

1.2. Metodología _____ **7**

1.3. Estructura _____ **7**

2. PROPUESTA DE TRADUCCIÓN _____ *9*

3. TRADUCCIÓN COMENTADA. Problemas de traducción _____ *22*

3.1. Sintagmas nominales complejos _____ **22**

3.1.1. Ejemplos de sintagmas nominales complejos y solución de los mismos _____ 22

3.1.2. Conclusión del apartado de sintagmas nominales complejos _____ 30

3.2. Adjetivos en inglés terminados en –al _____ **31**

3.2.1 Ejemplos de adjetivos en inglés que terminan en –al _____ 31

3.2.1.1. Adjetivos de relación o relacionales y su traducción _____ 32

3.2.1.2. Adjetivos restrictivos y su traducción _____ 33

3.3. Falsos amigos y léxico de la lengua general que constituyen un problema de traducción _____ **34**

3.3.1. Ejemplos por orden de aparición en el texto original: _____ 35

4. CONCLUSIONES _____ *42*

BIBLIOGRAFÍA _____ *47*

ANEXO 1 – Texto Original: "Integrating Genomes" _____ *49*

ANEXO 2 – Fig. 1. Ensamblaje y alineación _____ *59*

ANEXO 3 – Fig. 2. Los procesos dinámicos _____ *60*

ANEXO 4 – Captura de pantalla de RefWorks _____ *61*

ANEXO 5 – Captura de pantalla de archivos de DéjàVu y Trados _____ *62*

1. INTRODUCCIÓN

El presente trabajo académico consiste en la traducción comentada de un artículo científico enmarcado en la disciplina de la genética. El proyecto de traducción corresponde al ámbito de la traducción científica y tiene una relación intrínseca con las actividades de traducción vinculadas a uno de los perfiles profesionales que he cursado durante los dos últimos años de la carrera. Considero que la elección de la temática de este trabajo también está determinada por el interés y la necesidad personal de continuar con el aprendizaje y desarrollo de ciertas habilidades y competencias traductoras necesarias e inherentes a la traducción médica. De igual manera, el proceso de traducción y el análisis riguroso —tanto de las unidades léxicas problemáticas como de otros aspectos concretos en los que se centra este trabajo— me ayudarán a perfeccionar la tarea traductora con vistas a futuras traducciones y desde la perspectiva de un encargo de traducción real.

La formación continua del traductor especializado, sobre todo del traductor médico, es indispensable dado el carácter multidisciplinario de la traducción científica. Aunque es cierto que no es imprescindible ser un profesional de una disciplina científica para traducir un texto especializado de dicha disciplina, pienso que sí es necesario que el traductor posea conocimientos sólidos de las disciplinas que traduce con el fin de realizar una traducción *adecuada*. En mi opinión, solo se puede obtener un buen resultado si se combina el conocimiento lingüístico —sobre todo de la lengua meta— con el conocimiento especializado de la temática del texto. El conocimiento lingüístico por si solo puede dar como resultado un texto con términos incorrectos, ambigüedades o errores de contenido. No obstante, poseer solo el conocimiento científico tampoco es suficiente, dado que para transmitir de forma unívoca el significado del texto original también es fundamental expresarlo correctamente.

Por todo lo antes expuesto, he decidido continuar estudiando para formarme como traductora especializada en el campo de la biomedicina, especialmente en el campo de la genética —y, concretamente, en el del genoma humano—. Por lo tanto, la selección del ámbito del artículo científico sobre el cual realizo la traducción comentada viene determinada por el interés en este campo, el deseo de adquirir más conocimientos sobre los conceptos de esta disciplina y, en un futuro, dedicarme a la traducción de textos

especializados. No obstante, se debe tener en cuenta que la disciplina de la genética es muy amplia y abarca campos científicos tales como el genoma humano, la clonación y los transgénicos. Asimismo, con respecto a la disciplina en la que se contextualiza mi trabajo, considero que es de gran interés para la sociedad, puesto que, gracias a los avances en este campo se pueden identificar nuevos genes y determinar su contenido, desarrollar mapas genéticos, conocer las causas de muchas enfermedades, y diseñar y mejorar el diagnóstico, la prevención y la terapia de diferentes tipos de enfermedades y trastornos genéticos. Sin duda, la evolución de la genética implica el incremento de la esperanza y calidad de vida de las personas.

El artículo a traducir se titula "Integrating Genomes" (Zerbino, 2012), y es un texto actual publicado el 13 de abril del año en curso. Dicho artículo es una revisión de la prestigiosa revista científica *Science* AAAS, que pertenece a la Asociación Americana para el Avance de la Ciencia (en inglés, *American Association for the Advancement of Science, AAAS*). He seleccionado un artículo de esta revista porque elige con seriedad y exhaustividad los artículos que publica. Los manuscritos pasan por un proceso de arbitraje complejo por parte del comité del Cuerpo Editor de *Science* —que está compuesto por numerosos científicos— para ser validados posteriormente de acuerdo a su originalidad y rigor científico. Los artículos arbitrados son sometidos luego a otros dos niveles de análisis antes de ser publicados. En todos los procesos se tienen en cuenta —entre otros parámetros— la calidad de la investigación, la estructura del artículo y la ausencia de errores. Todos estos elementos los he valorado y tenido en cuenta a la hora de seleccionar el artículo para este trabajo académico.

Finalmente, me gustaría mencionar que realizar los estudios de Traducción e Interpretación era un reto personal y algo que deseaba hacer desde hace muchos años. Aunque en un inicio me gustaba más la traducción jurídico-económica, a medida que avanzaban los estudios descubrí más aptitud e inclinación hacia la traducción biomédica. Reconozco que aún tengo un largo camino de formación por recorrer para adquirir el máximo de conocimiento especializado posible y las competencias que requiere la traducción especializada en este ámbito. Sin embargo, el empeño y la dedicación constante que me caracterizan harán que llegue hasta la meta que me he trazado. De este modo, también

contribuiré con la sociedad, pues pienso que el traductor biomédico desempeña un papel importante en la transmisión del conocimiento científico y de los avances de la medicina.

1.1 Objetivos

Los objetivos de este trabajo académico consisten, en primer lugar, en realizar una traducción de calidad profesional en el contexto de un encargo real de traducción, reflexionar sobre las características de los textos médicos en general y exponer las que más se evidencian en el artículo a traducir y que constituyen un problema de traducción. Asimismo, y mediante la propuesta de traducción, comentar las diferentes estrategias utilizadas en los problemas tratados para que el texto meta sea lo más adecuado posible de conformidad con la tipología textual a la que pertenece. Y, por último, adquirir conocimientos sobre la disciplina de la genética y algunas subdisciplinas afines con la finalidad de traducir de forma eficiente artículos de este ámbito en el futuro.

1.2. Metodología

En primer lugar, se presentará la propuesta de traducción de un artículo de *Science*. Posteriormente, se expondrán por apartados tres de los problemas de traducción que se detectaron en el texto original: sintagmas nominales largos o complejos, adjetivos terminados en –al y falsos amigos y léxico de la lengua general. En cada apartado se mostrará cada ejemplo de forma individual —la mayoría de ellos incluirá el contexto— conjuntamente con la solución proporcionada en el texto meta, la estrategia utilizada en función del problema y, dado el caso, la justificación y otras posibles soluciones derivadas de la fase previa de documentación. Asimismo, y según el apartado, los ejemplos contemplarán las fuentes bibliográficas de referencia. Por último, en el apartado de conclusiones se resumirán los principales problemas de traducción tratados y el cumplimiento de los objetivos, junto con otras cuestiones relevantes.

1.3. Estructura

El trabajo comenzará con la propuesta de traducción. A continuación, se expondrán los tres apartados de problemas de traducción. Mayoritariamente, cada apartado estará compuesto por una breve introducción del tema, los ejemplos puntuales a tratar según el problema, las estrategias de traducción, la propuesta de traducción y, por último, la

conclusión del tema. Los apartados aparecerán de acuerdo al orden en que son tratados y se formularán tal y como se explica a continuación.

En primer lugar, a nivel sintáctico, los sintagmas nominales largos o complejos. Se expondrán ejemplos de sintagmas nominales complejos de más de 3 palabras (*tokens*) detectados en el texto original, junto con la solución proporcionada en la propuesta de traducción y los patrones sintácticos de ambos sintagmas. Además, los ejemplos contemplarán una breve explicación de los elementos que lo componen, su posición dentro del sintagma y otras explicaciones relevantes según el caso. A nivel morfológico, se mostrarán los adjetivos terminados en –al, la propuesta de traducción, así como la explicación y tratamiento de los mismos en la lengua de llegada. Por último, a nivel léxico, se tratarán algunos falsos amigos y léxico de la lengua general que son problemáticos a la hora de traducir, acompañados de las propuestas de traducción, una breve explicación de la ambigüedad o dificultad que presentan y las fuentes bibliográficas donde se han documentado.

Posteriormente, se presentará la conclusión del trabajo académico que contemplará un breve resumen en correspondencia con los objetivos, así como también una síntesis de las aportaciones, los argumentos y las conclusiones de los análisis realizados. Por último, se incluirá la bibliografía del trabajo y los anexos necesarios, comenzando por el texto original y las figuras 1 y 2 correspondientes al mismo (con la traducción de los elementos incluidos en ambas en un tamaño más grande para una visualización más clara en caso de dudas).

2. PROPUESTA DE TRADUCCIÓN

REVISIÓN

La integración de los genomas

Zerbino, D. R., Paten, B., & Haussler, D.

RESUMEN

Los proyectos de secuenciación del genoma intentan una integración cada vez más significativa de la información genética, molecular y fenotípica, motivo por el que ha surgido una nueva especialización de la genómica enmarcada en la subdisciplina de la genómica computacional. Los modelos heredados de la genética de poblaciones, la filogenética y las enfermedades genéticas humanas se fusionan con los de la teoría de grafos, la estadística, el procesamiento de señales y la informática para proporcionar una base cuantitativa sólida en el ámbito de la genómica que solo puede llevarse a cabo con la ayuda de un ordenador. Los datos obtenidos —derivados del rápido incremento de las muestras del 10^{30} de los organismos del planeta— tendrán un impacto en los diferentes campos de la ciencia y, a la vez, proporcionarán una perspectiva nueva y extraordinaria de la historia de la vida.

Los ordenadores han sido necesarios para procesar (*3*) y archivar (*2,4*) secuencias de datos desde el momento en que se obtuvieron las primeras secuencias de genomas a mediados de la década de los setenta (*1,2*). Sin embargo, la genómica computacional tiene su origen en el año 1980, cuando Smith y Waterman desarrollaron un algoritmo para encontrar con rapidez la comparación óptima (alineamiento) de dos secuencias de longitud n entre las más de 3^n posibilidades (*2,5*) y Stormo *et al.* crearon una función umbral lineal para buscar los sitios de unión de los ribosomas en una colección de 78 000 nucleótidos de una secuencia de ARNm de *E. coli*. Lo que antes parecía grandes conjuntos de datos para el ámbito de la biología en la actualidad no lo son, puesto que las máquinas de secuenciación de alto rendimiento de lectura corta generan terabytes de datos (*2,7*). En los últimos 8 años se ha observado una mejoría del rendimiento de secuenciación de 10 mil veces, muy superior a la mejoría estimada, en 16 veces, según el poder computacional de la ley de Moore (*8*). Actualmente, se considera que el uso de datos genómicos para establecer el modelo de la

evolución, el mecanismo y la función del genoma es el núcleo de un campo dinámico. Cada genoma es el resultado de una trayectoria evolutiva —mayoritariamente compartida, pero en parte singular— de 3 800 millones de años desde el origen de la vida. La diversidad se obtiene principalmente por errores de copia del ADN producidos durante el proceso de replicación. Estos errores producen cambios en una sola base, los cuales se conocen como sustitución si se diseminan a toda la población (fijación) o polimorfismos de nucleótidos simples (del inglés SNPs) si se presentan de manera desigual en la población (segregación). Los errores de replicación también producen inserciones y deleciones (*indels* es el acrónimo de ambos), así como duplicaciones en tándem cuando una secuencia corta se repite de forma consecutiva. Los cromosomas a menudo intercambian largos segmentos de ADN similares mediante el proceso de recombinación homóloga. Existen secuencias específicas de ADN conocidas como elementos transponibles (transposones) que tienen la capacidad de replicarse en el interior de la célula utilizando mecanismos análogos a los que se encuentran en determinados virus, y dejan muchas copias del transposón (*9*). Los reordenamientos producen patrones tales como las inversiones, las deleciones y duplicaciones de segmentos (que causan las variantes del número de copias, en inglés CNVs), así como la fusión y fisión de los cromosomas y las translocaciones cromosómicas (*10*). A gran escala, a veces se duplica el genoma completo, lo cual aumenta considerablemente el contenido génico del mismo (*11*). La diversidad de vida actual se obtuvo de forma gradual por medio de las ediciones y se manifiesta en el genotipo de línea germinal de todos los seres vivos. Comenzando por el genotipo de línea germinal, los genomas de las células somáticas padecen ediciones similares durante el periodo de vida de cada individuo y algunos experimentan un tipo de proceso evolutivo denominado selección somática, la cual desempeña un papel determinado en el cáncer y en la inmunidad (*2,12*).

Los genomas son el núcleo de los mecanismos moleculares de las células y de las propiedades físicas (o fenotipo) de los organismos. Además, contienen la información de las moléculas activas de la célula, o sea, las proteínas y el ARNm, así como otros ARN funcionales. La tecnología de secuenciación se utiliza para determinar la abundancia de ARN (*13*), la localización subcelular (*14*), las isoformas de corte y de empalme (*15*), la estructura secundaria (*16*) y el coeficiente de relación con complejos moleculares tales como el ribosoma (*17*). Dicha tecnología también se utiliza para evaluar los mecanismos epigenéticos que regulan la producción y la función del ARN y de las proteínas, entre los

que se incluye la metilación (*2, 18*), las modificaciones de histonas, la unión de factores de transcripción (*20*), la accesibilidad de la cromatina (*21*) y las interacciones tridimensionales de la cromatina (*2, 22*). La genómica computacional, cuando se aplican estos datos, produce modelos de mecanismos epigenéticos y redes reguladoras de genes (*2, 23*) que se articulan con modelos más amplios de biología de sistemas moleculares tales como las cascadas de señalización de las proteínas, las vías metabólicas y los motivos de redes reguladoras (*24*).

La genómica computacional interpreta los datos genómicos en tres dimensiones mediante la combinación de modelos evolutivos, mecanicistas, y funcionales. Un gen es, simultáneamente, una secuencia de ADN que evoluciona en el tiempo (historia), un fragmento de cromatina que interactúa con otras moléculas (mecanismo) y, como producto génico, un elemento de las vías de actividad en la célula que afecta al organismo (función). Las primeras fases del recorrido del genotipo hasta el fenotipo del organismo, y donde actúa la selección evolutiva, son los fenotipos moleculares procedentes del estado epigenético y de los niveles de expresión del ARN. Los tres desafíos fundamentales e interrelacionados de la genómica computacional —más allá de los conceptos básicos de almacenamiento, indexación y búsqueda de genomas de todas las especies vivas del mundo— son explicar la evolución del genoma, establecer modelos de los fenotipos moleculares como resultado del genotipo y predecir el fenotipo del organismo.

La obtención de las secuencias del genoma

Los métodos actuales para los análisis del genoma comienzan con el ensamblaje del genoma (*2*), que es el proceso de reconstrucción de un genoma completo a partir de fragmentos aleatorios de ADN relativamente cortos denominados secuencias. Es posible detectar las secuencias solapadas y, de ese modo, reconstruir progresivamente la mayor parte de las secuencias del genoma dado que existe suficiente redundancia de secuencias o profundidad de la cobertura de secuenciación (*2*). Sin embargo, este contexto ideal se complica por el hecho de que los genomas contienen habitualmente largas regiones redundantes (repeticiones de terminales largas, en inglés LTR) o regiones en las que la distribución estadística de las bases es altamente sesgada (regiones de baja complejidad), lo que da lugar a secuencias solapadas por coincidencia o falsas. Este hecho produce complejas redes de solapamientos en las secuencias, pero no todas reflejan un solapamiento

real en el genoma. La dificultad que más persiste en el ensamblaje es determinar cuáles son los solapamientos reales y cuáles son falsos. Este es un problema NP–complejo, lo que significa que es al menos tan difícil como cualquier problema del tipo que se resuelve en una máquina de Turing no determinista en tiempo polinómico (*2*). Por este motivo, se espera que las únicas soluciones eficaces sean los métodos heurísticos, los cuales no garantizan que se pueda encontrar la solución óptima. Por esta razón, las regiones complejas de los genomas se catalogan como espacios indeterminados (*2*), propensos a errores (*2*), o costosos de terminar. El hecho de que las tecnologías más actuales pueden generar lecturas más largas puede que ayude a paliar este problema.

Por lo general, después de ensamblar el primer genoma completo de una especie (genoma de referencia), no se vuelven a ensamblar los genomas procedentes de esas especies o de otras estrechamente relacionadas a ella. Dichas especies se ensamblan entonces utilizando el genoma de referencia como patrón, y se explotan las semejanzas que derivan del antepasado evolutivo común. Las secuencias del nuevo genoma se sitúan (alinean) frente al genoma de referencia (*2*) y se detectan las discordancias sistemáticas (*2*). Este proceso se puede utilizar simplemente para enumerar las variantes presentes en el nuevo genoma o para guiar el ensamblado completo del nuevo genoma (denominado ensamblado de referencia, en inglés reference-base assembly) (*2*). Sin embargo, las secuencias cortas —que suelen ser altamente repetitivas— pueden confundir a los algoritmos de alineamiento (2), razón por la que los ensambladores que funcionan tanto por referencia como por información cartográfica tienden a sesgarse hacia el genoma de referencia y, a menudo, interpretan a las variantes reales como ruido (*2*). Se espera que se produzcan mejoras a este respecto a medida que la tecnología avanza para obtener secuencias más largas.

La modelización evolutiva del genotipo

Los genomas se comparan por medio de las alineaciones: las bases de ADN se fragmentan en dos conjuntos (columnas) que proceden presuntamente de un ancestro común reciente idóneo. A partir de este proceso, se puede analizar lo que se ha conservado o cambiado durante la evolución de los genomas con respecto a su ancestro común. A gran escala, las alineaciones pueden indicar cambios tanto en el orden de los segmentos como en el número de copias, y, en una escala menor, pueden indicar sustituciones de base específicas (véase la Fig. 1. ampliada en el anexo 2 del presente trabajo).

Fig. 1. Ensamblaje y alineación. **(A)** Ensamblaje de varias secuencias, agrupadas en parejas de secuencias solapadas. Dado que la secuencia genómica contiene una secuencia repetida, las lecturas procedentes de las dos copias de la secuencia solapada debe separarse mediante un programa de ensamblaje de secuencias para originar un ensamblaje lineal. **(B)** Alineación de cinco secuencias y un grupo externo. Cada fila es una secuencia y cada columna es un par de bases que proviene de una base ancestral común. Hay seis columnas señaladas, de ellas, la columna cuatro contiene una base fija en las cinco secuencias, mientras que las otras columnas contienen SNPs segregados. Los árboles que se encuentran a cada lado indican dos maneras alternativas de representar la filogenia entre las secuencias. El árbol de la izquierda es óptimo en términos de complejidad de sustitución para las columnas 1, 2 y 3, mientras que el de la derecha es óptimo para las columnas 4, 5 y 6. Dada la diferencia entre los dos árboles, en las columnas 3 y 5 puede haber tenido lugar un evento de recombinación

El reto principal, tal y como ocurre en el ensamblaje de genomas, es distinguir las similitudes de las secuencias falsas de las que pertenecen a la ascendencia común. Las regiones de genomas que son objeto de la selección purificadora en la que se conserva la similitud de la secuencia —como las regiones ortólogas codificantes de proteínas— a menudo pueden alinearse de forma fiable a lo largo de grandes distancias evolutivas, tal y como ocurre entre los vertebrados y los invertebrados. Las regiones que se encuentran en una posición neutral (es decir, que no están bajo selección positiva o negativa), tienden a separarse mucho más rápido y pueden alinearse de forma fiable solo en el caso de que hayan divergido recientemente (p. ej. dos genomas de vertebrados que hayan divergido durante los últimos 100 millones de años) (*2*). Por lo tanto, es habitual distinguir las alineaciones de subregiones (alineaciones locales) (*2*) de las alineaciones de secuencias completas (alineaciones globales) (*2*) e, incluso, de las de genomas completos (alineaciones del genoma completo) (*2*). Por lo general, las alineaciones locales se utilizan entre las regiones funcionales conservadas de genomas más distantes (2). En cambio, las

alineaciones de genomas completos son apropiadas cuando se comparan genomas de especies más cercanas.

Los métodos filogenéticos proporcionan una orden explícita del gen de ascendencia mediante el ancestro compartido cuando se aplican a más de dos especies o a múltiples copias de genes de una especie. Cuando el modelo de evolución queda restringido a tener en cuenta solo las inserciones y deleciones (indels) y las sustituciones —que son los sucesos más comunes— entonces, la filogenia se representa mediante un solo árbol en el que los nodos terminales (hojas, en inglés leaf) representan las secuencias observadas (actuales), las ramas indican las líneas directas de origen y los nodos internos las presuntas secuencias ancestrales (*2*). El hecho de encontrar el árbol filogenético óptimo según los modelos de probabilidad o de parsimonia de sustituciones (y también de *indels)* es un problema NP-complejo (*2*), y se ha dedicado un esfuerzo considerable a la obtención de soluciones heurísticas eficientes y exactas.

La recombinación homóloga hace que los análisis filogenéticos sean más complejos, lo que produce moléculas de ADN cuyos elementos tienen historias evolutivas diferentes (Fig. 2). El modelo coalescente con recombinación (*2*) se utiliza para modelar la historia evolutiva de un gen que ha padecido sustituciones o recombinación homóloga. Las historias individuales de los componentes se representan en un alineamiento con un árbol filogenético separado (coalescente) para cada base (*2*) (Véase a continuación la Fig. 1. también ampliada en el anexo 2 del presente trabajo).

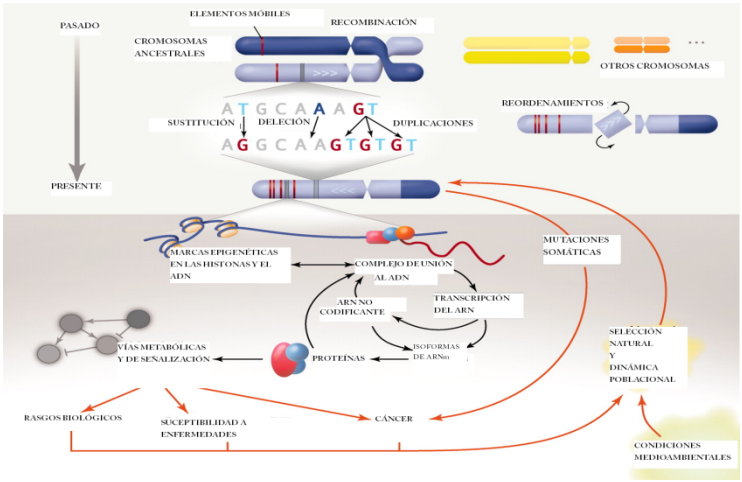


Fig. 2. Los procesos dinámicos que afectan al genoma y que, a su vez, son afectados por él. **Parte superior:** Las mutaciones aleatorias modifican al genoma y producen cambios en el mismo. A gran escala, los eventos de recombinación homóloga intercambian elementos equivalentes de ADN, los reordenamientos conectan diferentes regiones de ADN, y los transposones pueden reproducirse por sí solos. A menor escala, ocurren eventos en forma de pequeñas modificaciones como las sustituciones y las inserciones y deleciones. **Parte inferior:** El genoma influye en los procesos moleculares de la célula, concretamente en la transcripción de genes y en el ARN funcional, el cual afecta a través de las vías al fenotipo del organismo puesto que causa fenotipos como enfermedades y otros rasgos específicos. Los fenotipos, mediante la selección natural, condicionan la presión selectiva sobre el genoma que puede favorecer o no las mutaciones específicas

Las relaciones evolutivas entre las secuencias de ADN también pueden incluir reordenamientos estructurales equilibrados que cambian el orden y la orientación de las bases en el genoma, así como las duplicaciones de segmentos, las adiciones y las pérdidas que modifican el número de copias de bases homólogas (*2*). Desafortunadamente, estos procesos, por lo general, se diseñan y tratan de forma individual y separados también de las sustituciones e inserciones o deleciones cortas. Continúa siendo un reto importante para el campo la elaboración de una teoría unificada de modelos computacionales de base matemática y algorítmica para tratar la evolución del genoma, en la que los procesos estocásticos describan conjuntamente la sustitución, la recombinación y la reordenación de las bases, así como las diversas formas de duplicación, adición y pérdida (*2*). Es difícil predecir cuándo —si alguna vez ocurre— se podrán obtener estos modelos dado el conocimiento incompleto de las dificultades matemáticas inherente a tales modelos. De lo único que existe seguridad es de que proyectos como *1000 Genomes Project* (*2*) producirán grandes cantidades de datos a partir de los cuales se crearán y probarán diversos modelos aproximados mediante métodos de probabilidad.

Del genotipo al fenotipo

Los genetistas han relacionado las mutaciones genéticas con las diferencias fenotípicas durante muchos años, pero en la actualidad lo hacen a una escala sin precedentes. Los estudios de secuenciación de los vertebrados [Genoma 10K (*2*)], de los insectos [i5K (*2*)], de las plantas (*2*), de los microorganismos (*2*), de los estirpes celulares (*2*) y de los 'metagenomas' —realizados mediante la secuenciación de ADN procedente de muestras ambientales que contienen una colección desconocida de microorganismos (*2*)— revelan

decenas de miles de genomas y constituyen un reto para que se revisen y profundicen los métodos que se utilizan. Hasta la fecha, estos estudios han proporcionado ejemplos concretos del desarrollo de la historia y de la diversidad de la vida, han explorado los vínculos entre las poblaciones microbianas del organismo y la salud humana, y han estudiado la respuesta de las especies a los cambios medioambientales actuales tales como el cambio climático, las enfermedades y los competidores (*2*). Para realizar comparaciones más a fondo, se podrían poner en común estudios futuros con los datos experimentales procedentes de una expansión de recursos sobre cultivos celulares de diversas especies y tejidos (*2*), así como con las nuevas metodologías de análisis de organismos unicelulares (*2*).

Cuando se estudia la genética de poblaciones de una especie en concreto, la frecuencia de recombinación determina la probabilidad de que las variantes de las secuencias proximales compartan el mismo árbol coalescente (*2*). La falta de recombinación conlleva a un desequilibrio de ligamiento en el que se correlacionan las variantes de segregación cercanas. Este fenómeno se utiliza para asociar las variantes de segregación específicas con rasgos fenotípicos o enfermedades —por ejemplo, en estudios de asociación del genoma completo realizados con la tecnología de microchips o datos de secuenciaciones incompletas (*2*)—. Sin embargo, este mismo fenómeno limita la resolución de estos enfoques para encontrar la variante causal real. Los estudios de la asociación del genoma completo tampoco tienen en cuenta los patrones de segregación de alelos en familiares cercanos. Los estudios futuros de correlación del genotipo con el fenotipo mediante el uso genomas completos utilizarán cada vez más el contexto del genotipo en casos y controles relacionados o no relacionados con el fenotipo, en combinación con mejores predicciones de los posibles efectos de las variantes del genoma para identificar las variantes causales (*2*).

Grandes proyectos como *ENCODE* (*2*), *mod*ENCODE (*2*) y *Epigenomics Roadmap* (*2*) proporcionan datos sobre el epigenoma y el mecanismo de transcripción que son necesarios para crear modelos de fenotipos moleculares en los que interviene el estado epigenético, la expresión de ARN y los niveles de proteína inferidos que, además, requieren de herramientas de análisis especializadas. Los navegadores de genoma tales como *Ensembl* (*2*) *y UCSC Genome Browser* (*2*) proporcionan una visión general de estos datos, además de conocimientos de trasfondo y diversos resultados de modelización. Los modelos matemáticos de estos procesos contienen numerosas variables latentes (hidden variables)

—a menudo una para cada sitio del genoma— debido a que muchos elementos epigenéticos claves, la expresión de ARN y la producción de proteínas no pueden medirse directamente y, por lo tanto, deben ser inferidos. Los enfoques incluyen los modelos ocultos de Markov (*2*), el factor de grafos (*2*), las redes Bayesianas (*2*), y los campos aleatorios de Markov (*2*). El modelo de inferencia (estimación de parámetros) y la aplicación del modelo (cálculo de probabilidades condicionales y marginales) son tareas computacionales de cálculos a gran escala.

El genotipo determina el fenotipo por medio del estado epigenético, trascripcional y proteómico. Los métodos de clasificación y de regresión que se utilizan para predecir el fenotipo a partir del genotipo pueden beneficiarse de los cálculos de los estados intermedios como aportaciones adicionales o alternativas (*2*). Estos métodos incluyen modelos lineales generalizados, redes neuronales (*2*) y máquinas de vectores de soporte —estas últimas se prefieren, en parte, debido a la capacidad de hacer frente a espacios de características introducidas de muchas dimensiones (es decir, con muchas variables de medición)—. En la actualidad, se pueden obtener más beneficios de la predicción del fenotipo mediante la incorporación de conceptos biológicos para mejorar el espacio de características introducidas —por ejemplo, sustituyendo los niveles de transcripción inferidos o los niveles de actividad proteica inferidos por las mediciones brutas de la expresión génica (*2*)— que por la utilización de técnicas aún más sofisticadas como la de clasificación y regresión.

Las aplicaciones del futuro

La interpretación de la historia evolutiva compartida de la vida comienza por el almacenamiento, la indexación y la comparación de los genomas. Sin embargo, para llevarlo a cabo —cuando existe un segmento de consulta (*query*)— se necesitan herramientas con el fin de producir con rapidez segmentos de ADN relacionados evolutivamente según el modelo de la evolución del genoma. ¿Cómo se podrán fabricar estos segmentos a medida que el volumen de datos del genoma aumenta rápidamente de petabytes (10^{15} bytes) a exabytes (10^{18} bytes)? Una posibilidad podría ser el uso de la compresión diferencial basada en la trayectoria evolutiva inferida de los genomas, en la que se representa cada secuencia como un conjunto de diferencias con respecto al progenitor inferido (*2*). Este método podría permitir la creación de una nueva red de información genética compacta, rápida de

consultar y que refleje de forma directa el origen natural del grado de relación entre los genomas.

La genómica ha tenido un gran impacto en la medicina, y lo seguirá teniendo. Como resultado, se espera que avancen las opciones terapéuticas del cáncer, puesto que las modificaciones genómicas son el origen de casi todos los cánceres (*2*). Estos cambios genómicos ocurren aleatoriamente en las células somáticas del organismo como consecuencia de los impactos ambientales o como resultado de un tratamiento. Las subpoblaciones que poseen células cancerosas diferentes genéticamente tienden a expandirse y a competir (*2*). La secuenciación de una muestra de tejido benigno de un enfermo de cáncer revela el genoma de este individuo al nacer —es decir, la línea germinal del genoma—. La comparación de dicho genoma con el que se obtiene a partir de una biopsia del tumor, por lo tanto, revela las mutaciones que han ocurrido con posterioridad en las células cancerosas del enfermo. El seguimiento de los genomas tumorales con este procedimiento —desde el inicio de la enfermedad y durante cada etapa del tratamiento— se convertirá en la metodología a seguir e indicará las decisiones terapéuticas (*2*). Los cambios que son fácilmente detectables solo a través de métodos computacionales en el genotipo, el estado epigenético, el patrón de la expresión génica, y la estructura de la vía activada proporcionarán información crucial sobre el estado del tumor durante la fase inicial del crecimiento tumoral y durante la aparición de resistencia a la terapia (*2*). Asimismo, se determinarán mejor las variantes genómicas específicas de tumores recurrentes, así como los fenotipos moleculares intermedios que originan el cáncer y determinan la respuesta del paciente a la terapia (*2*), hecho que se traducirá en un diagnóstico y tratamiento de cáncer más individualizado (*2*).

Otros campos de la medicina también se beneficiarán de los métodos y resultados computacionales. Por ejemplo, las células inmunitarias (inmunocitos) experimentan mutaciones específicas mediante rondas de selección somática (2), que van acompañadas de cambios en el estado epigenético, en el patrón de la expresión génica y en la estructura de la vía activada. Para establecer el modelo de la función de los inmunocitos y relacionar la respuesta inmunitaria con el antígeno se llevará a cabo una secuenciación exhaustiva de los receptores de células T y de los anticuerpos de células B (también denominadas linfocitos B) (2), acompañada de mediciones del genoma completo de la variación genética, el estado

epigenético y el patrón de la expresión génica de los inmunocitos. Los datos genómicos de alto rendimiento se utilizarán para la elaboración de vacunas (incluyendo la inmunoterapia del cáncer) y para el tratamiento de enfermedades infecciosas (2), autoinmunitarias y sistemas inmunodeprimidos derivados de la quimioterapia, las transfusiones, los trasplantes y las terapias con células madres.

Las variantes genómicas, el estado epigenético y el patrón de expresión desempeñan un papel importante en las terapias con células madres y en las aplicaciones científicas básicas de las células madres que solo pueden distinguirse mediante el uso de las herramientas computacionales (2). Las células madres pluripotentes inducidas (en inglés iPS) y las células reprogramadas directamente de un linaje específico se obtienen a partir de células somáticas (2) que ya han padecido mutaciones somáticas y se han cultivado en condiciones tales que pueden ser seleccionadas para más mutaciones (2). Estas mutaciones pronto serán accesibles mediante el análisis del genoma completo. Las mediciones de las modificaciones epigenéticas y de la expresión génica confirmarán el estado pluripotencial o la categoría del linaje específico de las células reprogramadas y se podrá verificar que se ha borrado la memoria epigenética del tejido del cual procedía. La interpretación de estos datos será de gran importancia dado que cada lote de células reprogramadas mostrará algunas mutaciones genéticas inesperadas, cambios epigenéticos y diferencias de expresión en el genoma completo, algunas de las cuales pueden acarrear consecuencias. En resumen, el futuro de las investigaciones sobre el cáncer, la inmunología y las células madres comprende los tres retos fundamentales de la genómica computacional: la explicación evolutiva (somática), la modelización molecular del fenotipo y la predicción del fenotipo del organismo.

Además de otras aplicaciones médicas, otras perspectivas similares desempeñan un papel determinado en las aplicaciones de la genómica en una gran variedad de campos, tales como la agricultura (*2*) y el estudio de la prehistoria del ser humano. La creciente disponibilidad de datos conlleva al desarrollo de herramientas complejas de análisis multidimensionales que integran secuencias de ADN, alineaciones, árboles filogenéticos, listas de variantes, análisis funcionales y del perfil epigenómico, modificaciones del fenotipo, entre otros. De este modo, para hacer frente a los retos de obtener la máxima información de cada experimento de secuenciación, se deben tomar prestados los avances de una

variedad de campos de investigación y vincularlos con modelos matemáticos de base efectuados con métodos numéricos. Existe una tensión entre la exhaustividad de los modelos y la eficacia de los mismos. A medida que esto ocurre, se desarrolla un modelo exhaustivo, pero computacionalmente implementable, de la evolución del genoma y de las repercusiones funcionales en el organismo, todo ello enmarcado en la genómica computacional. Sin embargo, aún creemos necesaria una formulación que sea lo suficientemente simple y expresiva para comparar los modelos, almacenar los datos y proporcionar los resultados en grandes cantidades de exabytes. Se prevé que en un futuro próximo la genómica computacional facilite una mejor capacidad para explorar y explotar las estructuras de los genomas y los procesos subyacentes en el núcleo de la vida.

Bibliografía y Notas

1. W. Fiers et *al.*, *Nature* 260, 500 (1976).

2. Para una lista completa de referencias por tema, véase la tabla S1 en la información complementaria.

3. T. R. Gingeras, J. P. Milazzo, D. Sciaky, R. J. Roberts, Nucleic Acids Res. 7, 529 (1979).

4. G. H. Hamm, G. N. Cameron, Nucleic Acids Res. 14, 5 (1986).

5. T. F. Smith, M. S. Waterman, J. Mol. Biol. 147, 195 (1981).

6. G. D. Stormo, T. D. Schneider, L. Gold, A. Ehrenfeucht, Nucleic Acids Res. 10, 2997 (1982).

7. E. R. Mardis, Trends Genet. 24, 133 (2008).

8. L. D. Stein, Genome Biol. 11, 207 (2010).

9. C. Feschotte, Nat. Rev. Genet. 9, 397 (2008).

10. L. Feuk, A. R. Carson, S. W. Scherer, Nat. Rev. Genet. 7, 85 (2006).

11. P. Dehal, J. L. Boore, PLoS Biol. 3, e314 (2005).

12. L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley, Nat. Rev. Cancer 6, 924 (2006).

13. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Nat. Methods 5, 621 (2008).

14. J. C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, S. Wiemann, EMBO Rep. 1, 287 (2000).

15. C. Trapnell et *al.*, *Nature* 28, 511 (2010).

16. J. G. Underwood et *al.*, Nat. Methods 7, 995 (2010).

17. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* 324, 218 (2009).
18. A. L. Brunner et *al.*, Genome Res. 19, 1044 (2009).
19. T. S. Mikkelsen et *al.*, *Nature* 448, 553 (2007).
20. G. Robertson et *al.*, Nat. Methods 4, 651 (2007).
21. A. P. Boyle et *al.*, Cell 132, 311 (2008).
22. M. J. Fullwood et *al.*, *Nature* 462, 58 (2009).
23. P. J. Mitchell, R. Tjian, *Science* 245, 371 (1989).
24. U. Alon, Nat. Rev. Genet. 8, 450 (2007).

Agradecimientos: Agradecemos a D. A. Earl por el diseño de las figuras y a E. Green, M. Häussler, J. Ma, D. Earl, H. Zerbino, R. Kuhn, G. Hickey, T. Pringle, K. Pollard, A. Krogh, R. Shamir, M. Waterman y R. Durbin por las correcciones y comentarios.
Respaldados por el Instituto estadounidense *Howard Hughes Medical Institute* (D.H.), el Centro de Análisis de Datos para la Enciclopedia de elementos del ADN (por sus siglas en inglés, ENCODE) perteneciente al Instituto Nacional de Investigación del Genoma Humano [en inglés, *National Human Genome Research Institute Data Analysis Center for the Encyclopedia of DNA Elements grant U01 (B.P.)*], y la Asociación Americana de la Investigación del Cáncer [en inglés, *the American Association for Cancer Research* (*Stand Up To Cancer/An Integrated Approach to Targeting Breast Cancer Molecular Subtypes and Their Resistance Phenotypes*) (D.R.Z.)].

Información complementaria

www.sciencemag.org/cgi/content/full/336/6078/179/DC1

Table S1

10.1126/science.1216830

3. TRADUCCIÓN COMENTADA. Problemas de traducción

3.1. Sintagmas nominales complejos

"El discurso especializado es siempre más preciso, más conciso y más sistemático que el discurso no especializado" (Cabré, 2003, p.13). Por esta razón, los textos científicos presentan características propias tales como la frecuencia de sintagmas nominales complejos a lo largo de todo el texto para reflejar el conocimiento especializado. Una de las funciones de dichos sintagmas nominales es la de condensar las ideas y conceptos nuevos que surgen debido al avance constante de las investigaciones de las diferentes disciplinas científicas. El problema principal que plantean estos sintagmas en inglés es la correcta identificación del orden de los premodificadores cuando se traducen al español.

El texto original a partir del cual se realiza este trabajo contiene una gran cantidad de sintagmas nominales —de ello se puede deducir que es un texto altamente especializado—. Sin embargo, solo se analizarán en este trabajo los sintagmas nominales compuestos por más de tres tokens, que son los que tienen un grado de dificultad más elevado a la hora de traducir. Estos sintagmas —como es habitual en esta tipología textual— están formados por sustantivos, aunque "pueden incluir adjetivos o adverbios con la finalidad de conseguir precisión y claridad en la definición o explicación de conceptos nuevos" (Alcaraz, 2000, p. 63). Otro rasgo distintivo de estos sintagmas es que carecen de conectores o nexos, lo que dificulta aún más su interpretación. A continuación se ofrecen ejemplos de algunos de los sintagmas nominales complejos que se observan en el texto de partida y la solución de los mismos.

3.1.1. Ejemplos de sintagmas nominales complejos y solución de los mismos

En primer lugar, se presenta el sintagma nominal complejo en contexto y su traducción, luego se identifican los patrones sintácticos de ambos sintagmas y se denominan cada uno de los elementos que lo conforman. Posteriormente, se localiza el elemento principal o núcleo del sintagma y, a partir de este punto, se intenta organizar tanto textual como discursivamente el sintagma de forma que sea apropiado en la lengua meta y que transmita de forma unívoca el concepto del texto de origen. Se debe tener en cuenta que la mayoría de las veces es necesario el uso de preposiciones, artículos, determinantes, e incluso, incluir

palabras para expresar el mismo concepto o idea en español —en el lenguaje especializado en español no se condensan ni el conocimiento científico ni los conceptos de la misma forma que en inglés—.

Ejemplos:

Contexto: *Stormo et al. built a linear threshold function to search a library of 78,000 nucleotides of Escherichia coli messenger RNA sequence for ribosome binding sites.*

1. **Escherichia coli messenger RNA sequence**

Patrón: [N +N +N]

Solución: Secuencia de ARNm de E. coli

Patrón: [N + Prep + N + Prep + N]

En este ejemplo se ha tomado como una sola unidad tanto el nombre de la bacteria como el de ARN mensajero y se han traducido con sus respectivas abreviaciones. El primer nombre a la derecha es el núcleo y el resto de elementos son los constituyentes. El elemento inmediato al núcleo es *messenger ARN*.

En este sintagma los modificadores preceden el núcleo *sequence* y tienen una interpretación restrictiva calificativa, es decir, designan a un tipo de secuencia y a un tipo de bacteria específicos. Dicho sintagma se compone en español de derecha a izquierda consecutivamente.

Contexto: *What seemed large data sets for biology then don't seem so today, as high-throughput, short-read sequencing machines churn out terabytes of data.*

2. **high-throughput, short-read sequencing machines**

Patrón: [(Adj+N) +(Adj+N) + Adj + N]

Solución: máquinas de secuenciación de alto rendimiento de lectura corta

Patrón: [N+ Prep + N + Prep + Adj + N + Prep + N + Adj]

En esta frase, aunque el complemento modificador *high-throughput* está separado por una coma, también modifica al núcleo *machine* y, a su vez, al sintagma *sequencing machines*. En este caso, el sintagma se ordena siguiendo el orden derecha-izquierda-centro puesto que, de lo contrario, se podría entender que *high-throughput* modifica a *short-read* y no a *sequencing machine* como es el caso (máquinas de secuenciación de lectura corta de alto rendimiento). Podría darse el caso de esta confusión dado que también existe el sintagma de carácter fijo dentro de este ámbito *lecturas de alto*

rendimiento. Asimismo, se deduce que el sintagma *short-read*, unido con un guión, es el modificador inmediato del sintagma *secuencing machines*.

Contexto: […] *the process of reconstructing an entire genome from relatively short random DNA fragments, called reads.*

3. relatively short random DNA fragments

Patrón: Adv + Adj + Adj + N+N

Solución: fragmentos aleatorios de ADN relativamente cortos

Patrón: [N + Adj + Prep + N + Adv + Adj]

En este sintagma se pueden identificar varias combinaciones de palabras o locuciones fijas en este ámbito tales como: fragmentos de ADN, fragmentos relativamente cortos de ADN, fragmentos cortos de ADN y fragmentos aleatorios de ADN. Ahora bien, es necesario organizar este sintagma nominal de manera que la frase sea gramaticalmente correcta y tenga un significado unívoco de acuerdo con el contexto del texto meta. En este caso, se ha comenzado con el núcleo del sintagma *fragments*, pero no hay un orden consecutivo (derecha-izquierda). Los constituyentes descriptivos que forman este sintagma nominal complejo se pueden expresar de más de una manera en español. Sin embargo, el adverbio *relatively*, solo modifica a *short*. Por lo antes expuesto, sería correcto decir: *fragmentos de ADN aleatorios y relativamente cortos* pero no *fragmentes de ADN relativamente aleatorios y cortos*. Así, a pesar de existir más de una posibilidad para expresar este sintagma —con el objetivo de sintetizar, y teniendo en cuenta las diferentes locuciones fijas existentes en este ámbito ya mencionadas al inicio de este párrafo— se ha optado por la solución dada al inicio.

Contexto: *Regions of genomes that are subject to purifying selection in which similarity of sequence is conserved, such as orthologous protein-coding regions* […].

4. orthologous protein-coding regions

Patrón: [Adj + (N + Adj) + N]

Solución: regiones ortólogas codificantes de proteínas

Patrón: [N + Adj + V+ Prep + N]

En este ejemplo, después de identificar el núcleo *regions*, se observa un adjetivo y dos palabras unidas por un guion que forman un modificador compuesto (*protein-coding*).

Este tipo de modificadores compuestos normalmente preceden al nombre al que modifican y es una característica típica del lenguaje médico especializado en inglés para simplificar la cadena de elementos en un sintagma nominal. Después de este análisis, se llega a la conclusión de que el adjetivo *orthologous* es un adjetivo calificativo que modifica al nombre *regions* —y no a *protein*— no solo por lo antes expuesto sino porque en este caso delimita al nombre. Por ejemplo, las *regiones ortólogas* son un tipo de región que se diferencia de las *regiones parálogas*, porque a pesar de que ambas son regiones ricas en genes similares, las primeras no proceden de un mismo genoma sino de poblaciones y especies diferentes. Así, si se encuentra la misma región ortóloga en dos especies diferentes quiere decir que ambas tuvieron un antecesor común. Después de esta reflexión, también se ha analizado la posibilidad de que la terminación del adjetivo *orthologous* sea –ica/–ico —lo que daría como resultado el adjetivo *ortológica* y no *ortóloga*— de acuerdo con las normas de terminación de los adjetivos en español. Se ha comprobado que el adjetivo *ortológico* existe, pero, según la RAE, proviene del nombre *ortología*, que significa "arte de pronunciar correctamente y, en sentido más general, de hablar con propiedad" (RAE, 2001). Un tanto más interesante es el resultado de la búsqueda del significado del adjetivo *ortólogo* (ga), el cual, según la RAE, es "una persona versada en ortografía" (RAE, 2001). Por lo tanto, la RAE no proporciona una acepción con el significado adecuado para este contexto en particular. No obstante, estos términos se han documentado en textos paralelos del ámbito de la genética y son correctos.

Contexto: *The construction of a mathematically and algorithmically tractable unified theory of genome evolution* […].

5. mathematically and algorithmically tractable unified theory

Patrón: [Adv + Conj +Adv + Adj + Adj + N]

Solución: teoría unificada de modelos computacionales de base matemática y algorítmica

Patrón: [N + Adj + Prep + N + Adj + Prep + N + Adj + Conj + Adj]

En este caso, la secuencia se traduce comenzando por el sintagma *unified theory* —en el que *theory* es el núcleo del sintagma. A pesar de que esta frase en inglés es de fácil comprensión, para reexpresarla en español con completa claridad, y teniendo en cuenta el contexto global del texto, ha sido necesario transmitir el sentido con una

frase más elaborada que incluye más elementos modificadores que en el texto en inglés. Para comenzar, por un lado se ha de saber que un modelo matemático es la representación de fenómenos o sucesos reales. Dichos modelos son fáciles de aplicar a otras ciencias exactas como la ecología, pero no tanto a los procesos biológicos. Por otro lado, las herramientas para la secuenciación del genoma se basan tanto en los modelos matemáticos como en algoritmos. Se entiende entonces que sea necesaria una teoría unificada de ambos modelos para tratar los diferentes procesos de forma conjunta y obtener así mejores resultados en la integración y el estudio de la evolución de los genomas. El adjetivo *tractable* se entiende bien en el contexto, pero ha causado ciertos problemas —aunque bien es cierto que la frase podría simplemente quedar de la manera siguiente: *teoría unificada tratable con modelos matemáticos y algorítmicos*—. No obstante, se ha optado por la solución dada al inicio porque se considera que es más explicativa.

Contexto: […] *preferred in part because of their ability to cope with very high-dimensional input feature spaces* […]

6. high-dimensional input feature spaces

Patrón: [(Adj +Adj) + N + N + N]

Solución: espacios de características introducidas de muchas dimensiones

Patrón: [N + Prep + N + Adj + Prep + Adv + N]

Los elementos premodificadores que componen este sintagma nominal son descriptivos. Se traduce siguiendo el orden de los modificadores más cercanos al núcleo, pero es necesaria la interpretación de los atributos *high-dimensional*, los cuales han dado lugar a la frase posmodificadora en español: *de muchas dimensiones* —el adjetivo *high* en este contexto no significa *alto* en el sentido de altura, sino *que tiene muchas dimensiones*—. El elemento más complejo dentro del sintagma es *input* porque es difícil de encontrar la solución correcta si no se sabe con certeza a qué hace referencia la frase. Para resolver este problema ha sido necesaria una documentación a fondo de los conceptos que abarca y la consulta de las funciones de una máquina de vectores de soporte, la cual alinea los puntos de entrada a un espacio de características. De este modo, analizando la función de la máquina, utilizando textos paralelos y por el contexto de la frase, se ha podido llegar a la solución adecuada.

Contexto*: Recurring tumor-specific genomic variants and intermediate molecular phenotypes* […].

7. recurring tumor-specific genomic variants

Patrón: Adj + (N + Adj) + Adj + N

Solución: variantes genómicas específicas de tumores recurrentes

Patrón: [N + Adj + Adj + Prep + N + Adj]

En este sintagma el núcleo es *variants*, y lo acompaña el elemento premodificador inmediato *genomic*. Por un lado, tenemos un modificador complejo unido por guión que modifica directamente al sintagma *genomic variants*. Por otra parte, el adjetivo *recurring* que modifica solo al nombre *tumor*. En este caso se ha entendido por *recurring* según lo que se explica en la cuarta entrada del diccionario *Merriam–Webster*:

> [4: to occur again after an interval: occur time after time <the cancer recurred>], la cual se corresponde con la segunda entrada de la RAE: [*recurrente*: 2. adj. Que vuelve a ocurrir o a aparecer, especialmente después de un intervalo] (RAE, 2001)

que este adjetivo hace referencia a los tumores que vuelven a aparecer. Según se ha documentado en textos especializados paralelos, es correcto decir en español *tumores recurrentes*.

Contexto: *Induced pluripotent stem (iPS) cells and lineage-specific directly reprogrammed cells are made from somatic cells* […].

8. lineage-specific directly reprogrammed cells

Patrón: [(N + Adj) + Adv + Adj + N]

Solución: células reprogramadas directamente de un linaje específico

Patrón: [N + Adj + V + Prep + Art + N + Adj]

Identificado el núcleo del sintagma *cells*, que se encuentra modificado por el elemento inmediato *reprogrammed*, se observa un adverbio que indica forma o manera. En este caso, se ha traducido el adverbio *directly* por *directamente*, puesto que aunque se desaconseja con frecuencia el uso de adverbios, no es incorrecto en esta frase y no hay un uso abusivo de estos en el texto meta. También se puede mencionar a este respecto que dicho gerundio a menudo se puede traducir con la técnica de trasposición —o sea, cambiando la categoría gramatical— utilizando los verbos *obtener* o *proceder* o la frase relativa *que se obtienen a partir de*. En cuanto al modificador complejo *lineage-specific*, ya se ha mencionado que el uso del guion es una formula muy

utilizada en el lenguaje médico especializado en inglés para condensar los conceptos especializados.

Contexto: *The increasing availability of data is leading to the development of elaborate multidimensional analysis tools* […]

9. elaborate mutidimensional analysis tools

Patrón: [Adj + Adj + N + N]

Solución: herramientas complejas de análisis multidimensionales

Patrón: [N + Adj + Prep + N + Adj]

En este sintagma se comienza a partir del núcleo *tools*. De los dos adjetivos que conforman la cadena premodificadora del sintagma, se debe tener en cuenta que *elaborate* modifica directamente al núcleo. El adjetivo *elaborate* se ha traducido por *complejas* porque se entiende mejor y es más explicativo en este contexto —aunque en español, y de acuerdo con el nombre al que modifica (*herramientas*), tiene el mismo sentido que el adjetivo traducido literalmente (*elaboradas*)—. En este caso, *herramientas complejas* hace referencia a las herramientas que integran e incorporan muchas variantes y datos para obtener el máximo de resultados óptimos posibles. Cabe mencionar que la frase también es correcta si se antepone el adjetivo al sustantivo: *complejas* herramientas, puesto que los adjetivos graduales "admiten variación en el orden" (RAE; AALE, 2010, p. 256). Sin embargo, se ha tenido en cuenta que en español, y sobre todo en el lenguaje especializado, el adjetivo se suele posponer al nombre.

Contexto: *We have seen a 10,000-fold sequencing performance improvement in the past 8 years, […]*

10. a 10,000-fold sequencing performance improvement

Patrón: [Art + (Adj. num. + sufijo) + Adj + N + N]

Solución: una mejoría del rendimiento de secuenciación de 10 mil veces

Patrón: [Art + N + (Prep+ Art) + N + Prep + N + Prep + Adj. Num. + Adv]

En este sintagma se debe comenzar por el artículo indeterminado *a* y pasar posteriormente al núcleo *improvement*. Del núcleo, traducimos de derecha a izquierda consecutivamente para completar la frase. El problema de esta frase radica fundamentalmente en la traducción del modificador complejo *10,000-fold*, una locución que se utiliza con frecuencia en los textos especializados. Al sufijo *–fold*

siempre lo modifica un adjetivo numeral que le precede e indica que algo se multiplica o aumenta/disminuye la cantidad de veces que indica el número. En este contexto *rendimiento de secuenciación* es un sintagma fijo muy utilizado que hace referencia al hecho de secuenciar cada vez más genomas o secuenciar los cambios que se producen en ellos para conocer y estudiar las diferencias genéticas, desarrollar mapas genéticos, y estudiar la evolución de las especies, entre otras ventajas —todo ello con una tecnología más avanzada, en menos tiempo y con un coste menor—.

Contexto: […] *that are putatively derived from a suitable recent common ancestor*.

11. a suitably recent common ancestor

Patrón: [Art + Adj + Adj + Adj + N]

Solución: un ancestro común reciente idóneo

Patrón: [Art + N + Adj + Adj + Adj]

En este ejemplo, se comienza con el artículo indeterminado *a* que precede al núcleo *ancestor*. Luego se observan tres adjetivos que preceden al núcleo. El primero a la izquierda, *suitable*, puede significar apropiado, apto, adecuado, idóneo y se considera el más problemático para traducir. En este contexto, puede ser viable el uso de los adjetivos *idóneo* o *apropiado*. Por otra parte, si partimos del núcleo y traducimos de derecha a izquierda, se observa el sintagma *ancestro común reciente*, el cual es un sintagma fijo en el lenguaje especializado que se encuentra con frecuencia en este ámbito y, por lo tanto, no crea dificultad alguna para traducir.

Contexto: *It requires tools to rapidly produce evolutionarily related segments of DNA according to a model of genome evolution* […].

12. evolutionarily related segments of DNA

Patrón: [Adv + Adj + N + Prep + N]

Solución: segmentos de ADN relacionados evolutivamente

Patrón [N + Prep + N + Adj + Adv]

En este sintagma, el núcleo es *segments* y va acompañado del nombre *ADN* que califica directamente a *segments* y se obtiene el sintagma *segmentos de ADN*. El adjetivo *related*, en este caso, tiene el significado de *associated*, *connected*, y, dentro del contexto global, que tienen un punto en común en la evolución de la especie.

3.1.2. Conclusión del apartado de sintagmas nominales complejos

A modo de resumen, se llega a la conclusión de que en los textos científicos en español, el núcleo del sintagma nominal siempre va al inicio (a la izquierda). Por este motivo, es de vital importancia tanto el reconocimiento del núcleo como la delimitación del sintagma completo —especialmente los sintagmas más complejos—. Asimismo, se puede observar que en español casi nunca hay elementos que premodifican al núcleo, estos se utilizan más en el género literario. Como se puede observar después del análisis de estos sintagmas complejos, por regla general, los modificadores que preceden al núcleo del sintagma nominal en inglés, en español se posponen.

En relación a la estructura del sintagma, no siempre se sigue el mismo orden para formar el sintagma en español. Algunos sintagmas con muchos tokens son muy difíciles de analizar, pero es una característica propia de los textos científicos que se encuentra con regularidad en esta tipología textual. También se puede observar que, con frecuencia, en los grupos nominales en inglés compuestos por dos, tres o más nombres (N + N/ N + N + N) se utiliza la preposición *de* o *de* + *artículo* para separar tanto el núcleo de los modificadores como estos entre sí. Una muestra de ello es el primer ejemplo de sintagmas nominales complejos presentado en este apartado: *Escherichia coli messenger RNA sequence*, en español *secuencia* ***de*** *ARNm* ***de*** *E. coli.*

Finalmente, se puede decir que la cantidad de elementos que compone el sintagma nominal viene determinada, por lo general, por la dificultad del concepto que se quiere expresar. Asimismo, entre más elementos componen el sintagma, existe menos posibilidades de frecuencia del mismo. Los sintagmas nominales complejos constituyen un rasgo de la traducción especializada, la mayoría son lo suficientemente independientes como para que se puedan interpretar sin contexto y difícilmente se pueden encontrar en recursos terminológicos como bases de datos o diccionarios. De aquí la importancia de la correcta identificación e interpretación de los elementos que componen el sintagma, así como de la adecuada comprensión del texto original y de los conceptos que este transmite tanto a nivel microtextual como macrotextual.

3.2. Adjetivos en inglés terminados en –al

En los textos especializados —y los científicos en particular— se le debe prestar especial atención a los adjetivos puesto que estos desempeñan un papel importante en el discurso científico para describir y calificar fenómenos o añadir información, entre otras funciones discursivas importantes.

> "El adjetivo es una clase de palabras que modifica al sustantivo o que predica de él aportando muy variados significados. En un gran número de casos, el adjetivo denota propiedades o cualidades" (RAE y Comisión de Gramática, 2009, p. 905).

En este trabajo se prestará especial atención a los adjetivos en inglés que terminan en –al y a la formación de dichos adjetivos en español. Con frecuencia se utilizan en español muchos adjetivos con la terminación –al, que, dada las normas de formación de sufijos en español, bien podrían adoptar otra terminación y ser adaptados a nuestro idioma. Algunos autores los consideran calcos del inglés que se introducen y arraigan muy rápido en nuestra comunidad de hablantes. Este tipo de fenómeno se produce probablemente por el hecho de que la mayoría de los artículos científicos se escriben en inglés, y, por consiguiente, muchos conceptos nuevos y neologismos se introducen por primera vez en esta lengua. Sin embargo, es evidente que en el lenguaje especializado en español también existen muchos adjetivos con dicha terminación que son correctos tales como *prenatal*, *occipital*, *intestinal*, *cervical*, entre otros. Los ejemplos que se muestran a continuación son tomados del texto original del presente trabajo.

3.2.1 Ejemplos de adjetivos en inglés que terminan en –al

Por lo general, los adjetivos que en inglés terminan en –al tienen en español las terminaciones –ico (a), –iano (a), –ario, –o y –oso. Se debe prestar atención a algunos adjetivos como *fungal* > *fúng**ico*** o *vi**ral*** > *vír**ico*** que tienden a traducirse de manera errónea. Con respecto a este último, y probablemente debido a la presión del inglés, ya se acepta la variante *vir**al*** en la RAE.

Los adjetivos *statistical* y *mathematical* tienen en español la terminación –ico (a). Los adjetivos con esta terminación, además de la mayoría que tienen la terminación –ar, –al (como *computational* >*computacional*) son característicos de la *interpretación relacional*, es decir, son modificadores restrictivos y "establecen cierta relación con un ámbito o dominio

representado por el sustantivo del que se derivan o las propiedades de este" (RAE y AALE, 2010, p. 254).

3.2.1.1. Adjetivos de relación o relacionales y su traducción

1. **Ancestral > ancestral** (RAE, 2001)

- Perteneciente o relativo a los antepasados
- Tradicional o de origen remoto

2. **Computational > computacional** (RAE, 2001)

- perteneciente o relativo a la informática
- inform. Dicho de un estudio o de un proceso: Que se adapta a ser tratado mediante computadores.

La segunda acepción es la válida según el contexto del término en el texto original.

3. **Functional > funcional** (Hidalgo, 2004, p. 499)

- Relativo a una función

4. **Individual > individual** (Hidalgo, 2004, p. 542)

- Único (a); para una persona; personalizado/ individuo(a); una persona/ relativo a un individuo

Este adjetivo aparece dos veces en el texto: una vez se ha traducido por *individuo* y la otra por *individual.*

5. **Internal > interno (a)** (Hidalgo, 2004, p. 547)

- Relativo al interior; dentro del cuerpo.

6. **Local > local** (RAE, 2001)

- Perteneciente o relativo al lugar.
- Que solo afecta a una parte del cuerpo.

En el contexto del texto original hace referencia a la alineación en una subregión de la secuencia, a diferencia de la alineación global (alineación de toda la secuencia), por lo tanto, la segunda acepción es la válida en este caso.

7. **Marginal > marginal** (Hidalgo, 2004, p. 574)

- Relativo a los márgenes; que no es central/ leve; poco intenso o poco importante / dudoso (a); poco claro

8. **Mathematical > matemático (a)** (RAE, 2001)

- Exacto, preciso
- Perteneciente o relativo a las matemáticas

9. **Statistical > estadístico (a)** (RAE, 2001)

- Perteneciente o relativo a la estadística

10. **Structural >estructural**

- Relativo a la estructura (Hidalgo, 2004, p. 700)

Resumen: Los adjetivos relacionales tienen el mismo género y número que el nombre al que modifican. Además, no admiten adverbios de grado ya que "no denotan propiedades o cualidades en relación con las cuales se pudieran interpretar esas formas de cuantificación" (RAE y Comisión de Gramática, 2009, p. 917) y no se anteponen al nombre al que modifican. Por consiguiente, sería incorrecto decir *la distribución* ***muy*** *estadística* o ***computacional*** *genómica* —en el primer caso por el adverbio de grado y, en el segundo, porque no preceden al nombre al que modifican—.

Se debe tener en cuenta que algunos adjetivos relacionales también se pueden utilizar como adjetivos descriptivos en contextos diferentes cuando tienen relación descriptiva admiten adverbios de grado y, se pueden anteponer según el caso.

3.2.1.2. Adjetivos restrictivos y su traducción

Este tipo de adjetivos suele ocupar, aunque existen excepciones, una posición posnominal. Dichos adjetivos se denominan calificativos si expresan cualidades o propiedades. A continuación se muestran ejemplos de adjetivos restrictivos calificativos en el contexto del texto meta.

11. **Coincidental > casual**

- Happening completely by chance, without being planned (anónimo, 2003, p. 29)
- Que sucede por casualidad (RAE, 2001)

12. Conditional > condicional (anónimo, 2008, p. 133)

- Estar supeditado (a) o condicionado (a) a algo
- Que incluye o lleva consigo una condición o requisito.

13. Optimal /optimum > óptimo (a)

- El mejor; el estado más deseado. (Hidalgo, 2004, p. 610)
- Sumamente bueno, que no puede ser mejor (RAE, 2001)

14. Practical > práctico (a)

- El adjetivo *practical* es polisémico, puede tener varias acepciones en inglés: *real, effective*, *clear thinking*, *suitable or usuful for a situation*, *using your hands* (Anónimo, 2003, p. 1282)
- Según la RAE: relativo o perteneciente a la práctica; se dice de los conocimientos que enseñan el modo de hacer algo; experimentado, versado; que piensa o actúa ajustándose a la realidad; que comporta utilidad; aplicación de una idea o doctrina; contraste experimental de una doctrina" (RAE, 2001).

En el contexto del texto original el adjetivo *practical* hace referencia a que las alineaciones son *prácticas* en el sentido de que son *idóneas*, *apropiadas*, *útiles*, *convenientes*.

3.3. Falsos amigos y léxico de la lengua general que constituyen un problema de traducción

"La terminología juega un papel relevante en las características de un texto científico" (Cabré, 2003, p. 15). Sin embargo, las unidades terminológicas son más fáciles de traducir puesto que hay muchas fuentes y recursos de documentación en los diferentes ámbitos de la ciencia. Sin embargo, las palabras de la lengua general constituyen un problema de traducción más complejo de resolver. Sobre todo, porque en los textos científicos en inglés se utilizan muchas palabras del lenguaje común o registro coloquial como términos técnicos. Por lo tanto, hay dos cuestiones fundamentales a resolver cuando se traducen los textos de tipología científica al español: determinar los falsos amigos —sin descartar ni los adjetivos ni los adverbios que puedan tener en español una traducción engañosa— e identificar las unidades del lenguaje común que se utilizan como términos técnicos en un

texto especializado y traducirlas con la terminología adecuada en español. Por lo general, este proceso hace que se eleve el registro en el texto meta. Por todo lo explicado, a la hora de traducir se debe tener en cuenta que los textos especializados en "inglés tienen un nivel de formalidad más bajo que las lenguas romances" (Newmark, 1992, p.219).

3.3.1. Ejemplos por orden de aparición en el texto original:

1. **Mapped/mappers/mapping**

Según el Merriam Webster:

- transitivo
  1- to make a map
  2- to assign (as a set or element) in a mathematical or exact correspondence
  3- to locate (a gene) on a chromosome
- intransitive
  1- of a gene: to be located
  2- to be assigned in a relation or connection

Según el Diccionario (Real Academia Nacional de Medicina, 2012, p.1034)

- adj. *mapeado* = cartografiado (a)
- n. *mapeamiento/ mapeo* = *cartografía*
- verbo *mapear* = cartografiar

Conclusión: Considero que es un término polisémico pues hay que tener en cuenta que no siempre significa *cartografiar* o *mapear* —este último más utilizado en la variante americana, aunque cada vez más también se utiliza en la variante peninsular—. Muchas veces también significa *ubicar* o *localizar* —por ejemplo, una secuencia o fragmento de ADN— mientras que en otras ocasiones puede significar buscar la *equivalencia* o *correspondencia* de algo —sobre todo en el contexto de la genética—.

Contexto: *Reads from the new genome are* ***mapped*** *onto the reference genome* […]. En este ejemplo *mapped* no significa *cartografiar*, sino que las secuencias del nuevo genoma se *ubican* o *posicionan* en un lugar determinado, en este caso, frente al genoma de referencia para detectar las diferencias entre el genoma nuevo y el genoma de referencia.

2. **Biased**

Según el Gran Diccionario Oxford (Anónimo, 2003, p. 990)

- tendencioso, parcial
- estar predispuesto, estar en contra de algo
- inclinación, tendencia
- en estadística: margen de error, sesgo

Según el Diccionario **(**Real Academia Nacional de Medicina, 2012, p. 1476)

- sesgado (a): que contiene o comporta sesgo
- sesgo: diferencia entre la media de un estimador el valor real del parámetro que se pretende estimar
- error constante al medir una magnitud en condiciones idénticas: por defecto del instrumento de medida o por una tendencia del observador
- error en los resultados de un estudio debido a una recolección defectuosa de los datos

Contexto: […] *regions where the statistical distribution of bases is significantly* ***biased*** […]. En este contexto, el término *biased* está contextualizado en el ámbito de la estadística y, por tanto, lo adecuado sería utilizar el adjetivo ***sesgada***.

3. Unique

(Navarro, 2005, p. 1060)).

- En la mayoría de los casos, *unique* no significa *único* (en inglés: *single*, *only*, *sole*), sino "excepcional, singular, extraordinario, especial, poco común, sin igual, distintivo".

(RAE, 2001)

- Único

1.- solo y sin otro de su especie

2.- singular, extraordinario, excelente

Contexto: *Every genome is the result of a mostly shared, but partly* ***unique*** […]. En este contexto *unique* se utiliza en el sentido de ***único*** = ***singular***. Hace referencia a que los seres humanos compartimos ciertas características genéticas, pero que cada persona individualmente tiene características fenotípicas *singulares* que la hace a la vez diferente de las demás.

4. Critical

(RAE, 2001)

- 1. Crítico

Perteneciente o relativo a la crisis/ Se dice del estado, momento, punto, etc., en que esta se produce/ Dicho del tiempo, de un punto, de una ocasión, etc.: Más oportunos, o que deben aprovecharse o atenderse […].

- 2. Crucial:

- Dicho de una situación o de un momento: Crítico, decisivo.

- Importante: que importa, que tiene importancia

Contexto: […] *the interpretation of these data will be of* ***critical*** *importance.* En este contexto, como *crítical* modifica a *importance,* podemos decir que la interpretación de los datos es de una importancia *crucial.*

5. Single

(Navarro, 2005, p. 940)

- Único, solo, simple, sencillo, individual, aislado, suelto, singular
- A menudo se sustituyen en español por los prefijos *mono–*y *uni–*.

Contextos:

- […] is represented by a **single** tree in which the terminal > se representa mediante un ***solo*** árbol [o un árbol individual]
- *These create* ***single****-base changes* […] > cambios en una **sola** base
- […] ***single****-nucleotide polymorphisms* […] > polimorfismo de nucleótido **simple**
- […] *and newer* ***single****-cell assay methodologies* […] > metodologías de análisis de organismos **unicelulares**
- *When studying the population genetics of a* ***single*** *species* […] > de una especie en **concreto** [en **particular]**.

6. Major

(Navarro, 2005, p. 613)

- Mayor (en inglés: *larger*, *greater*, *bigger*)

(RAE, 2001)

- Principal, significativo, importante, grave, serio, (además de otras acepciones que no son relevantes en este contexto).
- Mayor: que excede algo en cantidad o calidad; dicho de una persona: que excede en edad a otra; **importante** […]

Contexto: […] *the various forms of duplication, gain, and loss, remains a* ***major*** *challenge for the field.* En este contexto *major challenge* se puede traducir por un *reto* ***importante***. En la mayoría de los casos *major* en inglés significa en español *principal*, *significativo*, *importante*.

7. Environmental

(Navarro, 2005, p. 345)

1- ambiental

2- medioambiental (relativo al medio ambiente)

(RAE, 2001)

- ambiental: perteneciente o relativo al medio ambiente
- medioambienal: perteneciente o relativo al medio ambiente

Contexto: En los ejemplos tomados del texto de partida se puede observar como el adjetivo *environmental* se traduce diferente según el contexto.

- […] *obtained by sequencing DNA from* ***environmental*** *samples containing an unknown collection of organisms* > […] obtenidos mediante la secuenciación de muestras ***ambientales*** que contienen una […]
- […]*from* ***environmental*** *impacts* [...] > impacto ***medioambiental***

8. Organisms

(Navarro, 2005, p. 729)

- 1- organismo
- 2- microorganismo

(RAE, 2001)

- organismo: Conjunto de órganos del cuerpo animal o vegetal y de las leyes por que se rige/ ser viviente […]

- microorganismo > microbio > Nombre genérico que designa los seres organizados solo visibles al microscopio; p. ej. las bacterias, los infusorios, las levaduras, etc.

Contextos:

- […] *a comprehensive but computable model of genome evolution and its functional repercussions on* ***organisms*** *is taking shape,* […] > hace referencia a ***organismo*** porque no habla de microorganismos solo, sino de todos los organismos vivientes, incluyendo la especie humana.
- *Unleashed on a rapidly increasing sample of the planet's* 10^{30} ***organisms*** […] > se debe traducir por ***organismos*** por la misma razón que el ejemplo anterior.
- *Genomes are the core of the molecular mechanisms of cells, and of the physical properties (or phenotype) of* ***organism****s.* > se debe traducir por ***organismos*** como en los ejemplos previos.
- […] *obtained by sequencing DNA from environmental samples containing an unknown collection of* ***organisms****.* > En este contexto lo tenemos que traducir por ***microorganismos***, puesto que hace referencia a las muestras de microorganismos que se obtuvieron del entorno ambiental. Por lo tanto, solo hace referencia a los microbios que solo son visibles por microscopio.

Discusión: Este término es polisémico puesto que en inglés *organisms* puede hacer referencia tanto al cuerpo humano o animal como a los microorganismos (cuando se refiere a algún microbio). En este último caso, a veces se puede precisar más y utilizar términos como *bacteria, virus, hongo*, según el contexto. Por lo tanto, hay que tener en cuenta el contexto para traducirlo correctamente.

9. Geneticist

El especialista en genética no se llama *geneticista*, sino *genetista* y, según la RAE, es la "persona que cultiva o domina los estudios de genética" (RAE, 2001). Algunos autores como Fernando Navarro consideran que "la forma correcta debería haber sido *genético*", aceptado por la RAE desde el 2001. Esta consideración está determinada por el hecho de que "los nombres de oficio derivados de ciencias que terminan en –ica, se forman con la terminación –ico" (Navarro, 2005, p. 412). Según la búsqueda en Google, se puede observar el uso

erróneo de *geneticista*, aunque los resultados de la búsqueda están muy por debajo con respecto a *genetista*.

10. Traits

Este término puede traducirse por *característica*, *rasgo*, *cualidad*, *peculiaridad* o *particularidad*. Hay que tener especial cuidado con su traducción y tener en cuenta que cuando se habla de *phenotype traits* u otros contextos del ámbito de la genética siempre se traduce por *rasgo*.

11. Cáncer (adjetivo)

Parece bastante evidente la traducción de este adjetivo en inglés, no obstante, el traductor debe distinguir entre *canceroso* (*sa*) y *cancerígeno* (*na*) en español.

Según la RAE:

Canceroso: Perteneciente o relativo al cáncer.

Cancerígeno: Que puede provocar cáncer.

Contexto: […] *subpopulations of genetically distinct* **cancer** cells *expand and compete (2).*

Por lo anteriormente expuesto, en este contexto se debe traducir el adjetivo *cancer* como: células ***cancerosas***.

12. Immune

En el contexto médico *immune* (inmune) significa "que no es atacable por ciertas enfermedades y no se debe utilizar como sinónimo de *inmunitario*" (Navarro, 2005, p.505). Aunque la RAE admite esta acepción en la entrada de *inmune*, los médicos y científicos distinguen la diferencia entre estos dos adjetivos, así pues, los traductores también deben mantener esa diferencia para garantizar la claridad y precisión del texto. Por consiguiente, se debe distinguir entre:

Inmune: que presenta inmunidad;

no atacable por ciertas enfermedades (RAE, 2001).

Inmunitario: relativo a la inmunidad (RAE, 2001).

Inmunológico: relativo a las ciencias que se ocupa del estudio de la inmunología (Navarro, 2005, p. 505).

Contextos:

- *For example,* **immune** *cells undergo specific mutations* [...].

Por ejemplo, las células *inmunitarias* experimentan mutaciones específicas […].

- […] *will be used to model immune cell function and correlate* **immune** response *with antigen*.

[…] relacionar la respuesta *inmunitaria* con el antígeno.

4. CONCLUSIONES

En el presente trabajo académico se ha intentado, en primer lugar, presentar una propuesta de traducción de calidad profesional en el contexto de un encargo real. Asimismo, a partir del texto de partida, se ha analizado —tanto de forma global en el ámbito de la traducción especializada como de manera más puntual en la traducción biomédica— algunas de las características que presenta esta tipología textual con el propósito de formular los problemas que presenta y exponer las estrategias de traducción y las soluciones propuestas. En general, el proceso de traducción ha servido para reflexionar sobre la traducción especializada y obtener como resultado final un texto meta *adecuado,* con la claridad, precisión y univocidad que requiere esta tipología textual. Sobre el artículo a traducir, cabe mencionar que pertenece al ámbito de la genética y que, aunque en un principio parecía un texto especializado de dificultad media, resultó ser un texto altamente técnico que integra conceptos de diferentes disciplinas y ámbitos científicos como la estadística, la genética computacional, la bioquímica y la biología molecular.

Como bien se ha podido demostrar a través de gran parte de los ejemplos analizados en este trabajo, las convenciones que se utilizan en el lenguaje especializado en inglés no se corresponden con las que se utilizan en nuestro idioma. Aunque tanto el texto original como la traducción son especializados, se observa entre sus diferencias —a nivel macrotextual— que el lenguaje especializado en español tiende a ser más formal. También se pudo constatar que los recursos estilísticos y lingüísticos utilizados en el lenguaje especializado varían de un idioma a otro. Un ejemplo claro de ello es la frecuencia en inglés de sintagmas nominales complejos —que se utilizan para condensar el conocimiento especializado y los conceptos nuevos— y el modo en que estos se forman y organizan de manera diferente en español. El tratamiento de dichos sintagmas constituye un tema clave de análisis en el presente trabajo.

Respecto al problema de traducción de los sintagmas nominales complejos, se ha podido comprobar que en español se necesita utilizar artículos, preposiciones, conjunciones e incluso añadir palabras para que la frase sea coherente, se entienda, tenga el mismo sentido semántico y cumpla con las convenciones formales de la lengua. Por lo general, los elementos de estos sintagmas en inglés premodifican al núcleo, en cambio, en español lo

posmodifican. La reflexión anterior implica que en inglés el núcleo de dichos sintagmas siempre se encuentra a la derecha del sintagma, mientras que en español casi siempre está a la izquierda —el sintagma se constituye, por lo general, comenzando por el núcleo y a continuación los diferentes elementos que lo modifican—. Contrariamente a lo que dice Vázquez-Ayora (1977, p. 123) sobre la manera de resolver estos sintagmas nominales complejos en inglés comenzando por la traducción del modificador más cercano al núcleo […] y continuar de allí con los que siguen hacia la izquierda, se ha podido comprobar, a través de los ejemplos analizados en este apartado, que seguir este orden no siempre proporciona la solución correcta.

El principal problema de traducción de los sintagmas complejos es precisamente la arbitrariedad del orden que deben seguir en español los elementos premodificadores del núcleo de estos sintagmas en inglés, junto con la falta de relaciones explícitas entre ellos —especialmente la relación semántica—. Aunque existe un gran número de autores que abordan —tanto desde la perspectiva lingüística y de análisis de corpus como traductológica— las características de los textos especializados, entre ellas los sintagmas nominales complejos como una peculiaridad de esta tipología, se considera que gran parte de los trabajos ya existentes son más teóricos que prácticos y que no hay suficiente bibliografía que trate este tema a fondo y con más rigor desde el contexto traductológico. La posibilidad de contar con más recursos y parámetros de estandarización para resolver los problemas de traducción relacionados con este tipo de sintagmas sería de gran ayuda y muy necesario para los traductores de textos especializados.

Para un traductor es útil —pero no suficiente— el análisis de patrones gramaticales o combinaciones sintácticas específicas, así como las reflexiones y conclusiones que se obtienen a partir de trabajos ya realizados desde la perspectiva de la lingüística de corpus. Por todo lo expuesto anteriormente, se considera que sería más conveniente y productivo, por ejemplo, la elaboración de recursos tales como glosarios o diccionarios bilingües de sintagmas nominales complejos por ámbitos o disciplinas concretas. Sin embargo, el inconveniente de estos recursos sería la actualización —no se encontrarían en ellos todos los sintagmas posibles— puesto que gran parte de las disciplinas científicas, como es el caso de la genética, se caracterizan por ser muy amplias y estar en constante evolución gracias a los avances y descubrimientos continuos. Es por eso por lo que los sintagmas

nominales complejos no son fijos —es decir, cada uno de ellos varía y es singular— ya que hacen referencia a un concepto o realidad determinada según la disciplina, el destinatario y el grado de especialización del texto. No obstante, se estima que la elaboración de este tipo de recursos sería viable y de gran utilidad para los traductores.

Otro de los problemas a tener en cuenta a la hora de traducir un texto especializado, es el uso de los adjetivos. Como se ha mencionado en el apartado correspondiente a este tema, los adjetivos son importantes en el discurso científico puesto que describen y califican fenómenos y añaden información, entre otras funciones discursivas importantes. Concretamente, en este trabajo se ha abordado el problema de los adjetivos con la terminación *–al* en inglés, pero los problemas de traducción relacionados con los adjetivos no se restringen a los que tienen esta terminación. En este sentido, es importante tener en cuenta cómo se forman los adjetivos en español y cuáles son sus terminaciones según el caso porque existe tendencia a utilizar calcos del inglés. Un ejemplo claro y puntual es la traducción errónea de *fungal* por *fungal* (en inglés), cuando lo correcto es *fúngico.* A través de la ejemplificación de algunos adjetivos presentes en el texto original, se han descrito algunos tipos de adjetivos con esta terminación que tienen una función determinada —adjetivos relacionales y restrictivos— y se ha llegado a la conclusión de que no todos se pueden anteponer al núcleo ni son graduables. Además, hay que tener en consideración la función semántica de cada adjetivo dentro de un sintagma nominal complejo, puesto que dicha función puede determinar el orden de los mismos con respecto al núcleo del sintagma.

El tercer y último problema de traducción que se ha tratado en este trabajo es el de los falsos amigos y el léxico de la lengua general que figuran precisamente en el texto original. Los problemas relacionados con este tema suelen ser de difícil resolución puesto que, como se ha indicado anteriormente, los artículos especializados en inglés suelen ser menos formales y, a menudo, palabras del lenguaje coloquial adquieren un valor técnico. En consecuencia, es fundamental estar alertas y reconocer este tipo de léxico en la fase de lectura previa del texto original. Se considera que los últimos años han sido fructíferos en cuanto a la elaboración de documentos y recursos que abordan el problema de los falsos amigos y el léxico complejo del lenguaje médico en general —por ejemplo, el *Diccionario crítico de dudas* de Fernando Navarro y los recursos y documentos que ofrecen asociaciones

tales como la Asociación Internacional de Traductores y Redactores de Medicina y Ciencias Afines (TREMÉDICA)—. Sin embargo, aún queda mucho por hacer a este respecto porque la lengua evoluciona constantemente y algunos de estos recursos —como el diccionario antes mencionado— han quedado algo obsoletos tanto por el uso de la lengua general como por la evolución del lenguaje médico.

Haciendo mención a los objetivos trazados al inicio de este trabajo, cabe mencionar que estos se han cumplido puesto que la propuesta de traducción presentada es una traducción de calidad profesional, que bien pudiera ser el resultado de un encargo de traducción real. De igual manera, a través de la exposición y análisis de tres de las características de los textos especializados que constituían a su vez un problema de traducción, se han dado propuestas de traducción *adecuadas* de acuerdo con el contexto y la tipología textual. Además, la fase previa de documentación ha servido para buscar información sobre una gran variedad de conceptos relacionados con la disciplina de la genética y otros campos afines, por lo que se han adquirido conocimientos sobre conceptos estadísticos, genéticos y de otras disciplinas científicas asociadas a la temática del presente trabajo.

En un plano más personal, este trabajo me ha ayudado a conocer el nivel en que me encuentro desde un punto de vista profesional y a identificar los puntos endebles que pudieran limitar mi futuro como traductora de textos especializados biomédicos. Del mismo modo, me ha servido para poner a prueba y mejorar las competencias que he adquirido durante los cuatro años de estudio tales como las competencias de documentación y procesamiento de la información en la fase previa —dada la necesidad de adquirir conocimientos sobre los conceptos del área temática—. Esta fase es de vital importancia puesto que nos ayuda a encontrar una solución óptima en el menor tiempo posible. Otras de las competencias puestas a prueba fueron: la competencia lingüística —tanto de interpretación del texto original como de reexpresión del texto meta—; la competencia comunicativa —teniendo en cuenta la temática, el lenguaje y el estilo discursivo del texto especializado—; la competencia traductora —que incluye la capacidad de transferencia, de análisis, de investigación, de razonamiento lógico, de organización y coherencia y el saber utilizar diferentes estrategias según el problema de traducción concreto.

También he podido desarrollar la competencia instrumental a través del uso de las herramientas de traducción asistida por ordenador (TAO). En este caso, he puesto a prueba los conocimientos de los programas DéjàVu y Trados. Una vez realizada la traducción, pensé que sería útil crear una memoria de traducción que facilitara en un futuro la traducción de textos de este ámbito y, a la vez, hacer un repaso de la utilización de estas herramientas TAO. Por este motivo, primero utilicé DéjàVu para alinear ambos textos y crear una memoria de traducción de la alineación en formato .tmx —que es el formato estándar de intercambio de memorias de traducción— y posteriormente Trados para pretraducir el texto original utilizando dicha memoria. En el apartado de Anexos incluiré una captura de pantalla en la que se muestran los documentos generados en los diferentes procesos.

A modo de conclusión, me gustaría mencionar que este trabajo académico ha sido un instrumento importante tanto para la aplicación y mejora de las diferentes competencias empleadas en el proceso de traducción como para la adquisición de conocimientos conceptuales de la temática abordada. El análisis de los diferentes problemas de traducción tratados han resultado útiles para la correcta toma de decisiones en el proceso de traducción y para conseguir una traducción *adecuada* de calidad profesional. La aportación del presente trabajo académico consiste, fundamentalmente, en su utilidad como obra de consulta y como texto paralelo —especialmente para traducciones de este ámbito—. De este modo, contribuirá a resolver dudas lingüísticas —de unidades terminológicas, expresiones y léxico general—, a advertir y observar cómo se pueden realizar algunos cambios sintácticos necesarios en los enunciados, y a incrementar los conocimientos sobre la temática en cuestión.

BIBLIOGRAFÍA

LIBROS

Alcaraz Varó, E. (2000). *El inglés profesional y académico*. Madrid: Alianza.

Cabré, M. T. (2003). El lenguaje científico desde la terminología. Dentro Gutiérrez, B., *Aproximaciones al lenguaje de la ciencia* (p. 19-52). Burgos: Fundación Instituto Castellano y Leonés de la Lengua.

Hurtado Albir, A. (2001). *Traducción y traductología: Introducción a la traductología*. Madrid: Cátedra.

Newmark, P. (1992). *Manual de traducción*. Madrid: Cátedra.

Real Academia Española. Comisión de Gramática. (2009). *Nueva gramática de la lengua española*. Madrid: Espasa-Calpe.

Real Academia Española & Asociación de Academias de la Lengua Española. (2010). *Nueva gramática de la lengua española: Manual*. Madrid: Espasa Libros.

Vázquez-Ayora. G. (1977). *Introducción a la traductología*. Washington: Georgetown University Press. Recuperado el 6 de junio de 2012 desde http://books.google.es/books?id=AvlxrH3vUNQC&printsec=frontcover&dq=vazquez+ayora&hl=ca&sa=X&ei=t6rYT_uDLqeI0AX9meWEBQ&ved=0CDUQuwUwAA#v=onepage&q=vazquez%20ayora&f=false

DICCIONARIOS

Diccionario Cambridge compact English-spanish, español- inglés para estudiantes de inglés. (2008). Cambridge: Cambridge University Press.

Gran diccionario oxford. (2003). Oxford: Oxford University Press.

Hidalgo-Simón, A. (2004). *Elsevier's dictionary of medicine: Spanish-english and english-spanish: By Ana Hidalgo Simón*. Ámsterdam: Elsevier.

Longman dictionary of contemporary english. (2003). (4th ed.). Harlow: Longman.

Navarro, F. A. (2005). *Diccionario crítico de dudas inglés-español de medicina* (2ª rev, mej y muy ampliada ed.). Madrid: McGraw-Hill Interamericana.

Real Academia Nacional de Medicina. (2012). *Diccionario de términos médicos*. Madrid: Médica Panamericana.

ARTÍCULO (TEXTO META)

Zerbino, D. R., Paten, B., & Haussler, D. (2012). Integrating genomes. *Science*: *336*(6078), 179-182. Recuperado el 14 de abril de 2012 desde http://www.sciencemag.org/content/336/6078/179.full

OTROS RECURSOS EN LÍNEA

Real Academia Española. (2001). *Diccionario de la lengua española* (22ª ed.). Consultado en http://www.rae.es/rae.html

ANEXO 1 – Texto Original: "Integrating Genomes"

REVIEW

Integrating Genomes

Zerbino, D. R., Paten, B., & Haussler, D.

ABSTRACT

As genomic sequencing projects attempt ever more ambitious integration of genetic, molecular, and phenotypic information, a specialization of genomics has emerged, embodied in the subdiscipline of computational genomics. Models inherited from population genetics, phylogenetics, and human disease genetics merge with those from graph theory, statistics, signal processing, and computer science to provide a rich quantitative foundation for genomics that can only be realized with the aid of a computer. Unleashed on a rapidly increasing sample of the planet's 10^{30} organisms, these analyses will have an impact on diverse fields of science while providing an extraordinary new window into the story of life.

Since the first genome sequences were obtained in the mid-1970s (*1*, *2*), computers have been necessary for processing (*3*) and archiving (*2*, *4*) sequence data. However, the discipline of computational genomics traces its roots to 1980, when Smith and Waterman developed an algorithm to rapidly find the optimal comparison (alignment) of two sequences of length n among the more than 3^n possibilities (*2*, *5*), and Stormo *et al.* built a linear threshold function to search a library of 78,000 nucleotides of *Escherichia coli* messenger RNA sequence for ribosome binding sites (*6*). What seemed large data sets for biology then don't seem so today, as high-throughput, short-read sequencing machines churn out terabytes of data (*2*, *7*). We have seen a 10,000-fold sequencing performance improvement in the past 8 years, far outpacing the estimated 16-fold improvement in computational power under Moore's law (*8*). Using genomics data to model genome evolution, mechanism, and function is now the heart of a lively field.

Every genome is the result of a mostly shared, but partly unique, 3.8-billion-year evolutionary journey from the origin of life. Diversity is created mostly by copy errors during replication. These create single-base changes, which are known as substitutions if spread to the whole population (fixed) or single-nucleotide polymorphisms (SNPs) if not

uniformly present in the population (segregating). Replication errors also create insertions and deletions (collectively, indels), as well as tandem duplications where a short sequence is repeated sequentially. Chromosomes often exchange long similar segments through the process of homologous recombination. Specific sequences of DNA, known as transposable elements, have the capacity to replicate themselves within the cell, using machinery analogous to that found in certain viruses, leaving many copies (*9*). Rearrangements lead to patterns such as inversions, segmental deletions and duplications (causing copy number variants), chromosome fusion and fission, and translocations between chromosomes (*10*). At the largest scale, occasionally the whole genome is duplicated, greatly increasing its gene content (*11*). The present diversity of life was created gradually through these edits and is manifest in the germline genotype of each living individual. Starting from the germline genotype, the genomes of the somatic cells continue to experience similar edits during the lifetime of every individual, some undergoing a kind of evolutionary process called somatic selection that plays a role in cancer and immunity (*2*, *12*).

Genomes are the core of the molecular mechanisms of cells, and of the physical properties (or phenotype) of organisms. They contain recipes of the active molecules of the cell, proteins, and their messenger RNAs, as well as other functional RNAs. Sequencing technology is used to determine RNA abundance (*13*), subcellular location (*14*), splicing isoforms (*15*), secondary structure (*16*), and rates of engagement with molecular complexes such as the ribosome (*17*). It is used to assay the epigenetic mechanisms that regulate RNA and protein production and function, including methylation (*2*, *18*), histone modifications (*19*), transcription factor binding (*20*), chromatin accessibility (*21*), and chromatin three-dimensional interactions (*2*, *22*). When applied to these data, computational genomics builds models of epigenetic mechanisms and gene regulatory networks (*2*, *23*), articulating with the broader models of molecular systems biology such as protein signaling cascades, metabolic pathways, and regulatory network motifs (*24*).

Combining evolutionary, mechanistic, and functional models, computational genomics interprets genomic data along three dimensions. A gene is simultaneously a DNA sequence evolving in time (history), a piece of chromatin that interacts with other molecules (mechanism), and, as a gene product, an actor in pathways of activity within the cell that affect the organism (function). Molecular phenotypes from epigenetic state and RNA expression levels are the first stations on the road from genotype to organismal phenotype,

where evolutionary selection acts. Beyond the basics of storing, indexing, and searching the world's genomes, the three fundamental, interrelated challenges of computational genomics are to explain genome evolution, model molecular phenotypes as a consequence of genotype, and predict organismal phenotype.

Obtaining Genomic Sequences

Current methods in genome analysis start with genome assembly (*2*), the process of reconstructing an entire genome from relatively short random DNA fragments, called reads. Given sufficient read redundancy, or coverage depth, it is possible to detect read overlaps and thereby progressively reconstitute most of the genome sequence (*2*). However, this ideal scenario is complicated by the fact that genomes commonly contain large redundant regions (repeats), or regions where the statistical distribution of bases is significantly biased (low-complexity DNA), leading to coincidental, spurious read overlaps. These create complex networks of read-to-read overlaps that do not all reflect actual overlaps in the genome. The most persistent difficulty of assembly is to determine which overlaps are legitimate and which are spurious. This problem is NP-hard, which means that it is at least as hard as any problem in the class of problems that can be solved in nondeterministic polynomial time (*2*). Therefore, we expect that the only efficient solutions will be heuristic methods that are not guaranteed to find the optimal solution. For this reason, difficult regions of genomes are left as undetermined gaps (*2*), prone to errors (*2*), or costly to finish (*2*). Newer sequencing technologies, producing longer reads (*2*), may alleviate this problem.

After the first complete genome from a species is assembled (the reference genome), new genomes from that species or closely related species are generally not assembled de novo but are assembled using the reference genome as a template, exploiting similarities derived from the common evolutionary ancestor. Reads from the new genome are mapped (aligned) onto the reference genome (*2*), and systematic discordances are detected (*2*). This process may be used simply to enumerate the variants present in the new genome, or to guide the complete assembly of the new genome (called reference-based assembly) (*2*). However, with short reads, mapping algorithms may also be confounded by repeats (*2*), and both mappers and reference-based assemblers tend to bias toward the reference genome, occasionally treating genuine variants as noise (*2*). As technology improves toward longer reads, we expect improvements here too.

Modeling the Evolution of Genotype

Genomes are compared by alignment: The bases of DNA are partitioned into sets (columns) that are putatively derived from a suitably recent common ancestor. From this, we can analyze what was conserved or changed during the evolution of the genomes from their common ancestor. At a large scale, alignments can indicate changes in segment order and copy number, and at a small scale they can indicate specific base substitutions (see Fig. 1).

Fig. 1

Assembly and alignment. (**A**) Assembly of a number of reads, grouped by pairwise sequence overlap. Because the genomic sequence contains a repeated sequence, the reads coming from the two copies of the repeat overlap and must be separated by the assembly software to produce a linear assembly. (**B**) Alignment of five sequences and an outgroup. Each row is a sequence; each column is a set of bases that descend from a common ancestral base. Six columns are highlighted. Column 4 contains a base that is fixed among the five sequences, whereas the other columns contain segregating SNPs. The trees on the sides represent two alternative ways of representing the phylogeny between the sequences. The left tree is optimal in terms of substitution complexity for columns 1, 2, and 3; the right tree is optimal for columns 4, 5, and 6. Given the difference between the two trees, a recombination event may have occurred between columns 3 and 5.

As in genome assembly, the primary challenge is to distinguish spurious sequence similarities from those due to common ancestry. Regions of genomes that are subject to purifying selection in which similarity of sequence is conserved, such as orthologous protein-coding regions, can often be reliably aligned across great evolutionary distances, such as between vertebrates and invertebrates. Regions that are neutrally drifting (i.e., not under positive or negative selection) diverge much more quickly, and can be reliably aligned only if they diverged recently (e.g., within the past 100 million years for two vertebrate genomes) (*2*). It is therefore common to distinguish alignments of subregions (local alignments) (*2*) from alignments of complete sequences (global alignments) (*2*) or even

complete genomes (genome alignments) (2). Local alignments are typically used between conserved functional regions of more distantly related genomes (2). Conversely, full genome alignments become practical when comparing genomes from closely related species.

When applied to more than two species or to multiple gene copies within a species, phylogenetic methods provide an explicit order of gene descent through shared ancestry. When the model of evolution is restricted to consider only indels and substitutions (the most common events), the phylogeny is represented by a single tree in which the terminal (leaf) nodes represent the observed (present-day) sequences, the branches represent direct lines of descent, and the internal nodes represent the putative ancestral sequences (2). Finding the optimal phylogeny under probabilistic or parsimony models of substitutions (and also of indels) is NP-hard (2), and considerable effort has been devoted to obtaining efficient and accurate heuristic solutions.

Phylogenetic analysis is complicated by homologous recombination, which creates DNA molecules whose parts have different evolutionary histories (Fig. 2). The coalescent model with recombination (2) models the evolutionary history of a gene with both substitutions and homologous recombination. Individual histories of parts are represented in an alignment with a separate phylogenetic (coalescent) tree for each base (2).

Evolutionary relationships between DNA sequences may also include balanced structural rearrangements that change the order and the orientation of the bases in the genome, as well as segmental duplications, gains, and losses that alter the number of copies of homologous bases (2). Unfortunately, these processes are usually modeled and treated separately from one another, and separately from substitutions and short indels. The construction of a mathematically and algorithmically tractable unified theory of genome evolution, in which stochastic processes jointly describe base substitution, recombination, rearrangement, and the various forms of duplication, gain, and loss, remains a major challenge for the field (2). With incomplete knowledge of the mathematical difficulties inherent in such a model, it is hard to predict when, if ever, such a model will be forthcoming. The only thing we are assured of is that projects such as the 1000 Genomes Project (2) will be producing massive amounts of data from which to build and test (via likelihood methods) a variety of approximate models.

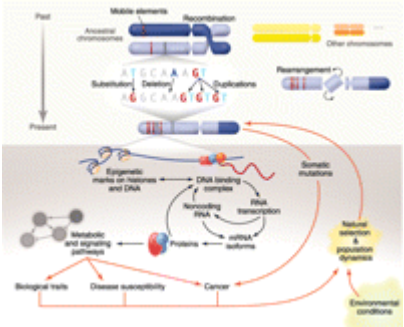


Fig. 2

The dynamic processes that affect and are affected by the genome. **Top**: The genome changes as it is modified by random mutations. At the larger scale, homologous recombination events swap equivalent pieces of DNA, rearrangements reconnect different regions of DNA, and transposable elements can self-reproduce. At the finer scale, small modifications such as substitutions and insertion/deletion events occur. **Bottom**: The genome affects the molecular processes in the cell, namely the transcription of genes and functional RNA, which through pathways affect the phenotype of the organism by causing phenotypes such as disease and other specific traits. Through natural selection, the phenotypes condition the selective pressure on the genome favoring or disfavoring specific mutations.

From Genotype to Phenotype

Geneticists have correlated genomic mutations to phenotypic differences for many years, but today they do so at an unprecedented scale. Sequencing surveys across vertebrates [Genome 10K (2)], insects [i5K (2)], plants (2), microorganisms (2), cell lineages (2), and "metagenomes" (obtained by sequencing DNA from environmental samples containing an unknown collection of organisms) (2) present us with tens of thousands of genomes and challenge us to rework and deepen our methods. To date, such studies have given us concrete examples of the unfolding history and diversity of life, explored the ties between the body's microbial populations and our health, and investigated the response of species to current environmental changes such as climate shift, disease, and competitors (2). Future studies could be coupled with experimental data derived from an expansion of cell culture resources for diverse species and tissues (2) and newer single-cell assay methodologies (2), allowing deeper comparisons.

When studying the population genetics of a single species, the recombination rate determines how likely it is that proximal sequence variants share the same coalescent tree (2). Lack of recombination leads to linkage disequilibrium, in which nearby segregating variants are correlated. This phenomenon is exploited in correlating specific segregating variants with phenotypic traits or diseases—for example, in genome-wide association studies conducted with microarrays or incomplete sequencing data (2). However, this same phenomenon limits the resolution of these approaches in finding the actual causal variant. Genome-wide association studies are also blind to the patterns of allele segregation in close relatives. Future genotype-phenotype studies using complete genomes will increasingly use genotypic context in related as well as unrelated cases and controls, combined with better prediction of the possible effects of genome variants, to identify causal variants (2).

Large projects such as ENCODE (2), modENCODE (2), and the Epigenomics Roadmap (2) are providing data on the epigenome and the transcriptional machinery needed to construct models of molecular phenotypes involving epigenetic state, RNA expression, and (inferred) protein levels, requiring specialized analysis tools. Genome browsers such as Ensembl (2) and the UCSC Genome Browser (2) provide an integrated view of these data, along with background knowledge and various modeling results. Because many key elements of epigenetics, RNA expression, and protein production cannot be directly measured and therefore must be inferred, the mathematical models of these processes contain numerous latent (hidden) variables, often one for every site in the genome. Approaches include hidden Markov models (2), factor graphs (2), Bayesian networks (2), and Markov random fields (2). Model inference (parameter estimation) and model application (computation of conditional and marginal probabilities) are large-scale computational tasks.

Genotype determines phenotype via epigenetic, transcriptional, and proteomic state. Classification and regression methods that are used to predict phenotype from genotype can take advantage of estimates of these intermediate states as additional or alternate inputs (2). These methods include general linear models, neural networks (2), and support vector machines (2), preferred in part because of their ability to cope with very high-dimensional input feature spaces (i.e., with many measured variables). There is currently more to be gained in predicting phenotype by incorporating biological knowledge to improve the input feature space—for example, by substituting inferred transcript levels or inferred protein

activity levels for raw gene expression measurements (2)—than by using yet more sophisticated techniques of classification and regression.

Looking Ahead to Applications

Understanding the shared evolutionary history of life starts by storing, indexing, and comparing genomes. It requires tools to rapidly produce evolutionarily related segments of DNA according to a model of genome evolution when prompted with a query segment. How will this be accomplished as we collectively grow from petabytes (10^{15} bytes) of genome data today to exabytes (10^{18} bytes) tomorrow? One possibility may be to use differential compression based on the inferred evolutionary trajectory of genomes, where each sequence is represented as a set of differences from its inferred parent (2). This may allow us to create a new web of genetic information that is compact, rapidly searchable, and directly reflects the natural origin of genomic relatedness.

Genomics has had a profound effect on medicine and will continue to do so. Cancer therapeutics are expected to advance as a result, because genomic modifications are the source of nearly all cancers (2). Within the body's somatic cells, genomic changes occur at random, from environmental impacts, or as a result of treatment; subpopulations of genetically distinct cancer cells expand and compete (2). Sequencing a sample of a cancer patient's noncancerous tissue reveals the patient's genome at birth (i.e., germline genome). Comparing this to the genome obtained from a tumor biopsy then reveals the mutations that have occurred subsequently in the patient's cancer cells. Tracking tumor genomes in this manner from early disease through each stage of treatment will become the norm and will inform therapeutic decisions (2). Changes that are readily detectable only through computational methods in genotype, epigenetic state, gene expression pattern, and activated pathway structure will provide crucial information on the state of the tumor during initial tumor growth and during the emergence of resistance to therapy (2). Recurring tumor-specific genomic variants and intermediate molecular phenotypes that drive cancer and determine patient response to therapy will come more clearly to light (2) and will be translated into better-targeted cancer diagnosis and treatment (2).

Other fields of medicine will also benefit from computational methods and findings. For example, immune cells undergo specific mutations through rounds of somatic selection (2), accompanied by changes in epigenetic state, gene expression pattern, and activated pathway structure. Deep sequencing of T cell receptors and B cell antibodies (2), coupled with

genome-wide measurements of genetic variation, epigenetic state, and gene expression pattern in immune cells, will be used to model immune cell function and correlate immune response with antigen. High-throughput genomics data will be used in vaccine design (including cancer immunotherapy) and the treatment of infectious diseases (2), autoimmune diseases, and compromised immune systems resulting from chemotherapy, transfusions, transplants, and stem cell therapies (2).

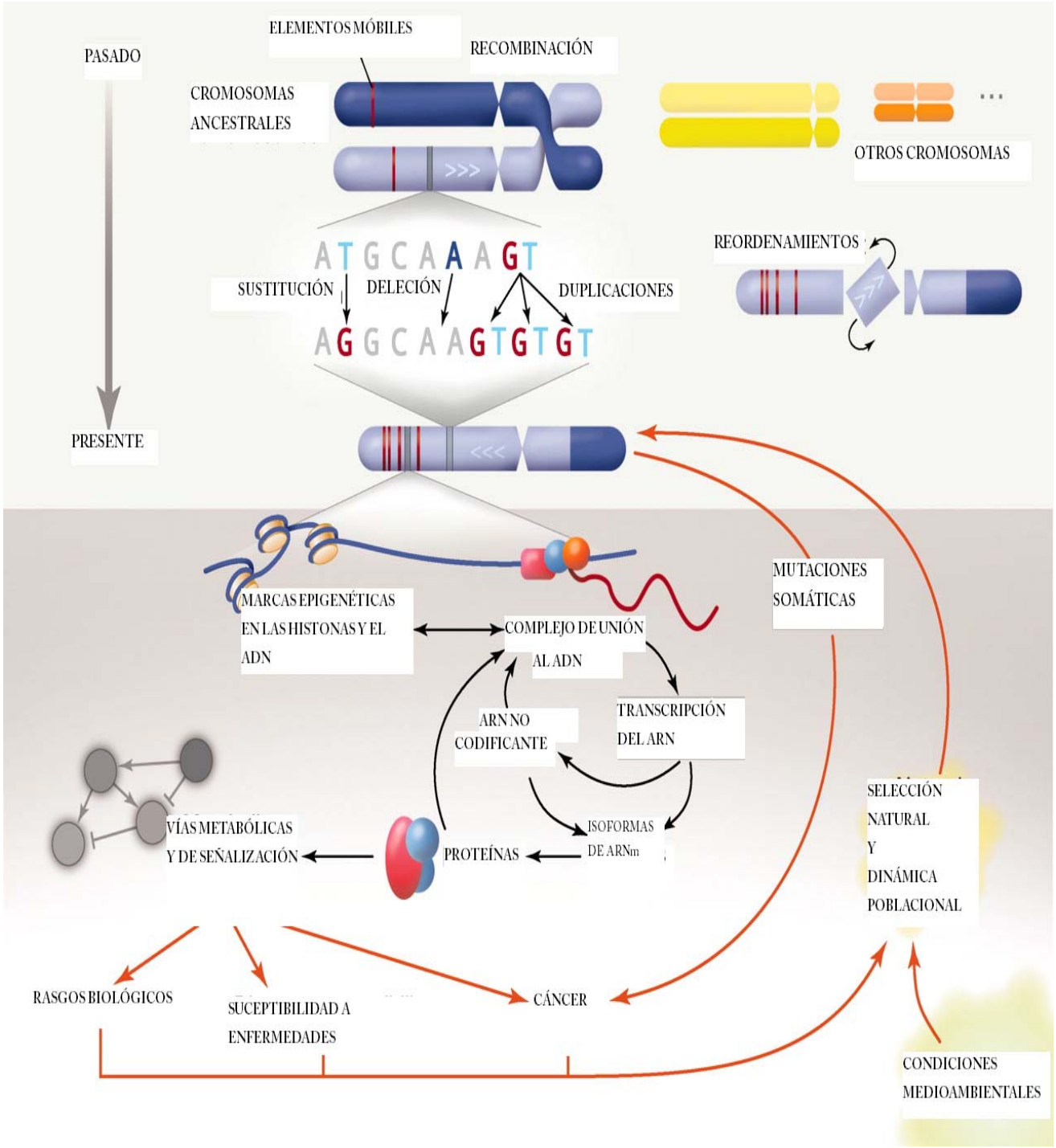
Genomic variants, epigenetic state, and expression pattern play key roles in stem cell therapies and basic science applications of stem cells that can only be discerned through the use of computational tools (2). Induced pluripotent stem (iPS) cells and lineage-specific directly reprogrammed cells are made from somatic cells (2) that have already incurred somatic mutations and are cultured in conditions that may select for further mutations (2). These mutations will soon be assessed with whole-genome analysis. Measurements of epigenetic modification and gene expression will confirm the pluripotent or lineage-specific status of the reprogrammed cells and verify that the epigenetic memory of the tissue from which they were derived is erased. Because every batch of reprogrammed cells will show some unexpected genetic mutations, epigenetic changes, and expression differences on a genome-wide level, some with consequences, the interpretation of these data will be of critical importance. In summary, the future of research into cancer, immunology, and stem cells involves all three key challenges of computational genomics: explaining (somatic) evolution, modeling molecular phenotype, and predicting organismal phenotype.

In addition to other medical applications, similar scenarios are playing out in applications of genomics in a wide range of fields, such as agriculture (2) and the study of human prehistory (2). The increasing availability of data is leading to the development of elaborate multidimensional analysis tools incorporating DNA sequences, alignments, phylogenetic trees, lists of variants, epigenomic and functional assays, phenotypic changes, etc. To face the challenges of obtaining the maximum information from every sequencing experiment, we must borrow advances from a spectrum of different research fields and tie them together into foundational mathematical models implemented with numerical methods. There is a tension between the comprehensiveness of models and their computational efficiency. As this plays out, a comprehensive but computable model of genome evolution and its functional repercussions on organisms is taking shape, embodied in computational genomics. Yet we still await a formulation that is both simple and expressive enough to

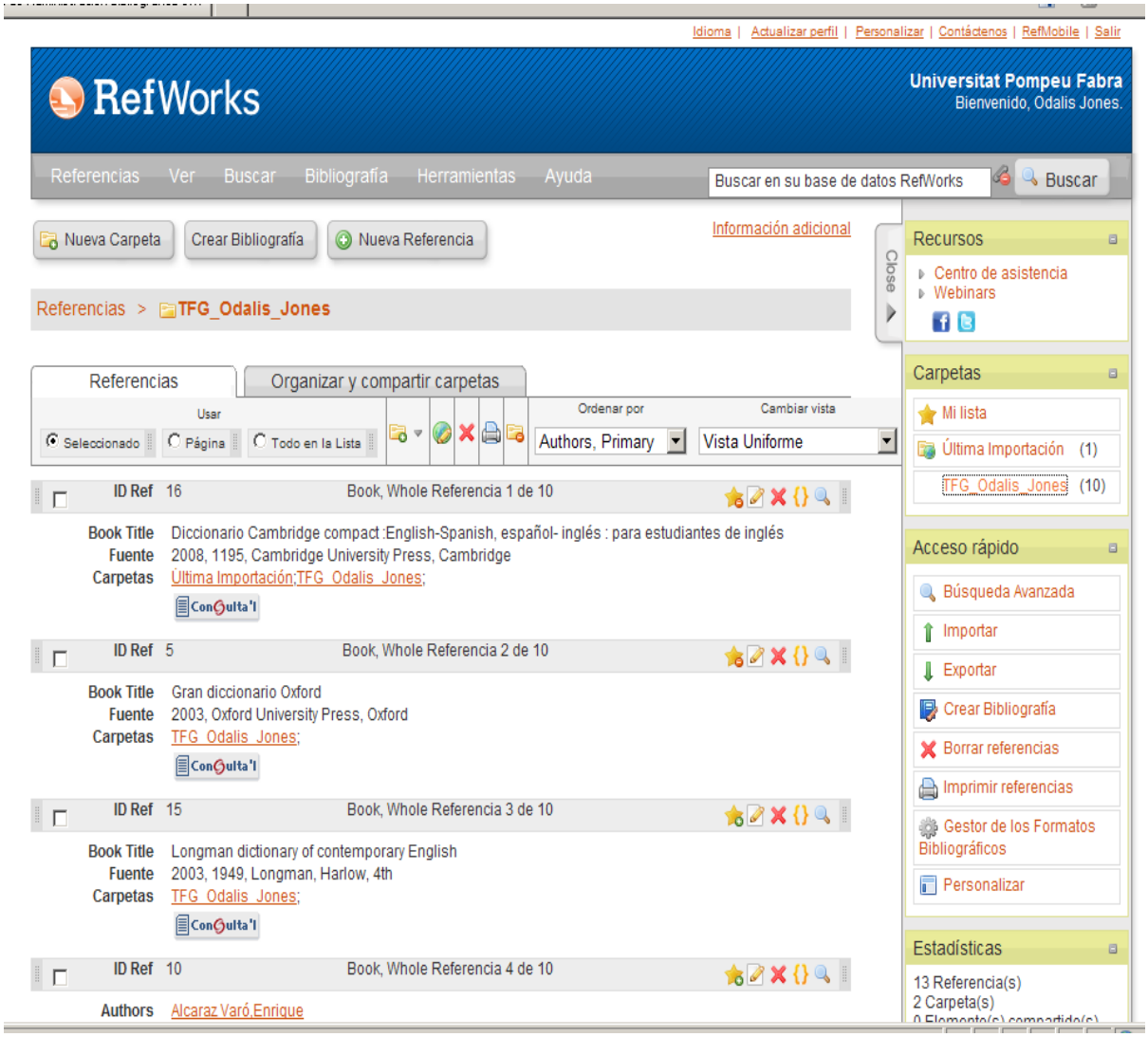
compare models, store information, and communicate results in an exabyte age. As a common language develops, shaped by our increasing knowledge of biology, we anticipate that computational genomics will provide enhanced ability to explore and exploit the genome structures and processes that lie at the heart of life.

ANEXO 2 – Fig. 1. Ensamblaje y alineación

ANEXO 3 – Fig. 2. Los procesos dinámicos

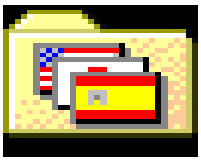


ANEXO 4 – Captura de pantalla de RefWorks



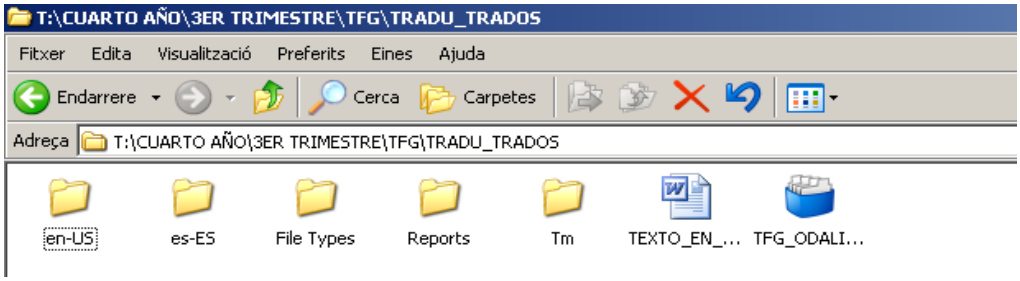
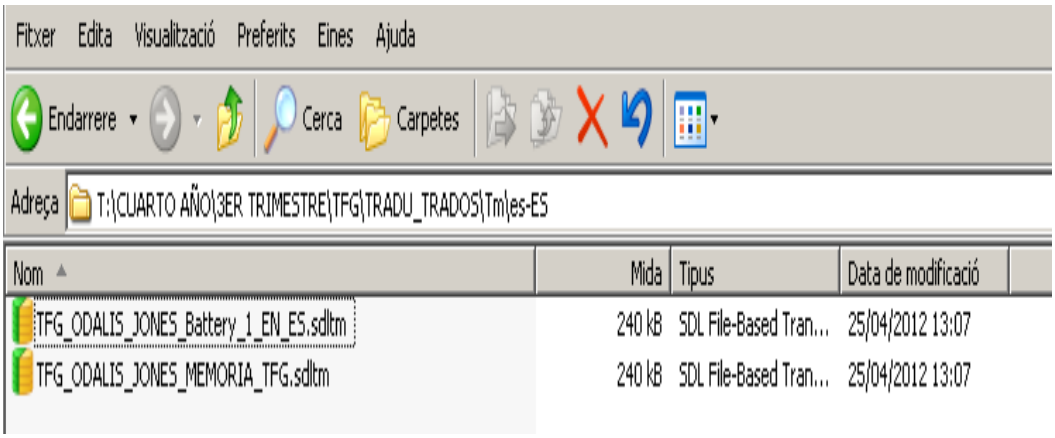
ANEXO 5 – Captura de pantalla de archivos de DéjàVu y Trados

Proyecto de alineación de TO y TM con DéjàVu



proy_tfg_dejav_ody.dvprj

Proyecto con Trados para crear memoria de traducción del TFG



Memoria de traducción