CGWork: Herramienta para el estudio de genómica comparativa con aplicaciones en el diseño racional de vacunas

CGWork: Tool for the study of comparative genomics with applications in the rational design of vaccines

Jorge Alejandro Jiménez Garí [1*], https://orcid.org/0000-0001-9586-5354
Fernando Figueredo Sánchez[2], https://orcid.org/0000-0003-4566-476X
Víctor Ventura Mena [3], https://orcid.org/0000-0003-1415-5604

[1] Departamento de Bioinformática, Facultad de Ciencias y Tecnológicas Computacionales, Universidad de las Ciencias Informáticas. jorgeajg@estudiantes.uci.cu
[2] Departamento de Bioinformática, Facultad de Ciencias y Tecnológicas Computacionales, Universidad de las Ciencias Informáticas. fernandofs@estudiantes.uci.cu
[3] Departamento de Bioinformática, Facultad de Ciencias y Tecnológicas Computacionales, Universidad de las Ciencias Informáticas. victorvm@estudiantes.uci.cu

*** Autor para correspondencia:** jorgeajg@estudiantes.uci.cu

Resumen

Debido al desarrollo de las técnicas de secuenciación y la acumulación de datos de varios genomas de microorganismos se abre la posibilidad de realizar estudios de genómica comparativa para la identificación de candidatos vacunales. En estos estudios se buscan proteínas de superficie universalmente presentes y poco variables entre organismos de una misma especie. En este trabajo se presenta una herramienta capaz de realizar un proceso de agrupamiento y análisis de subconjuntos de genomas. Basa su funcionamiento en un proceso de agrupamiento progresivo con dos medidas de similitud, una que permite el cálculo rápido con baja sensibilidad, basada en descomposición en k-tuplas, y otra que permite una búsqueda exhaustiva, basada en alineamientos, con una mayor sensibilidad, pero con mayor tiempo de procesamiento. Además, posibilita la búsqueda de un gen específico y su localización tanto en el genoma como los clústeres en que se encuentra, para la búsqueda de sus posibles ortólogos La herramienta tiene potencialidades en el diseño racional de vacunas, y ya se está empleando en un proyecto cuyo fin es la explicación de la reactividad cruzada a la vacuna cubana contra meningitis meningocócica VAMENGOC-BC.

Palabras clave: Genómica comparativa, análisis de pan-genoma, agrupamiento, diseño racional de vacunas.

Abstract

Due to the development of sequencing techniques and the accumulation of data from several genomes of microorganisms, the possibility of conducting comparative genomics studies for the identification of vaccine candidates opens up. These studies look for surface proteins that are universally present and not very variable between organisms of the same species. In this work, a tool capable of performing a process of grouping and analysis of subsets of genomes is presented. It bases its operation on a progressive grouping process with two similarity measures, one that allows rapid calculation with low sensitivity, based on k-tuple decomposition, and the other that allows an exhaustive search, based on alignments, with greater sensitivity. but with

longer processing time. In addition, it makes it possible to search for a specific gene and its location both in the genome and in the clusters in which it is found, in order to search for its possible orthologs. The tool has potential in the rational design of vaccines, and is already being used in a project whose purpose is the explanation of the cross-reactivity to the Cuban vaccine against meningococcal meningitis VAMENGOC-BC.

Keywords: *Comparative genomics, pan-genome analysis, clustering, rational vaccine design.*

Recibido: 20/12/2021
Aceptado: 18/03/2022
En línea: 01/06/2022

Introducción

Una secuencia genómica completa de un organismo puede considerarse como el mapa genético definitivo, en el sentido de que las características hereditarias están codificadas dentro del ADN y que se conoce el orden de todos los nucleótidos a lo largo de cada cromosoma (Hardison 2003). Sin embargo, el conocimiento de la secuencia de ADN no dice directamente cómo esta información genética conduce a los rasgos y comportamientos observables (fenotipos) que se quieren entender. Un gran reto para la siguiente fase de la investigación genómica es distinguir el ADN funcional y luego asignarle un papel (Miller, Makova et al. 2004). La genómica comparativa es uno de los principales enfoques utilizados en la anotación funcional de los genomas.

Esta rama del conocimiento estudia las semejanzas y diferencias entre genomas de diferentes organismos. Las secuencias funcionales están sujetas a selección evolutiva, que puede dejar una firma en las secuencias alineadas (Miller, Makova et al. 2004). Este enfoque busca beneficiarse de la información proporcionada por las firmas de la selección natural para entender la función y los procesos evolutivos que actúan sobre los genomas.
La predicción de genes es una aplicación importante de la genómica comparativa. Identificar genes homólogos posibilita el reconocimiento de la función desconocida de un gen, siguiendo el principio de que la homología implica posiblemente la existencia de funciones comunes, ya sea gracias a la determinación de genes ortólogos (genes homólogos presentes en distintos organismos que codifican proteínas con la misma función y que han evolucionado mediante descendencia directa) o de los genes parálogos (genes homólogos presentes en un mismo organismo que codifican proteínas con un origen común, que son el resultado de un proceso de duplicación en el genoma).

La secuenciación de alto rendimiento es ahora lo suficientemente rápida y barata como para ser considerada parte de la caja de herramientas para investigar las bacterias. Las secuencias del genoma bacteriano ahora pueden ser generadas internamente en muchos laboratorios, en cuestión de horas o días (Edwards and Holt 2013). Esto ha

generado una avalancha de datos disponibles públicamente en bases de datos de Internet. La explotación de las secuencias del genoma bacteriano ha proporcionado hasta ahora una gran cantidad de nueva información general sobre la diversidad genética de las bacterias, entre ellas muchos patógenos. La genómica comparativa ha permitido descubrir muchas variaciones genómicas incluyendo bacterias relacionadas y reveló principios básicos involucrados en la diversificación bacteriana, mejorando nuestro conocimiento de la evolución de los patógenos bacterianos (Dobrindt and Hacker 2001).

La subsiguiente secuenciación masiva de numerosos y completos genomas microbianos reveló nuevos fenómenos evolutivos, entre los cuales los más fundamentales son: (1) la transferencia horizontal generalizada de genes (THG), en gran parte mediada por virus y plásmidos, que da forma a los genomas de las arqueas y bacterias y exige una revisión radical (si no el abandono) del concepto de Árbol de la Vida, (2) la herencia de tipo lamarckiano que parece ser crítica para la defensa antivírica y otras formas de adaptación en los procariotas, y (3) la evolución de los mecanismos dedicados a la evolución, tales como los vehículos para los sistemas de THG y los sistemas de mutagénesis inducida por el estrés (Koonin and Wolf 2012). Este comportamiento ha llevado a desarrollar el concepto de plasticidad del genoma, que enfatiza la importancia de la adquisición y pérdida de genes para la evolución del genoma y el hecho de que la organización genética refleja el estilo de vida bacteriano (Dobrindt and Hacker 2001).

Dada esta facilidad para generar borradores de secuencias genéticas completas, sería útil saber cuántos genomas deben secuenciarse para que una especie determinada represente con precisión todo su repertorio genético. El término "pan-genoma" o "supragenoma" denota el conjunto de todos los genes presentes en los genomas de los miembros de un grupo de organismos, generalmente una especie (Lapierre and Gogarten 2009). El análisis pan-genómico, por el cual se caracteriza el tamaño del repertorio genético accesible a cualquier especie dada, junto con una estimación del número de secuencias genómicas completas necesarias para un análisis adecuado, se está aplicando cada vez más. Existen diferentes modelos para el análisis y su precisión y aplicabilidad dependen del caso en cuestión (Tettelin, Riley et al. 2008).

Para este fin se han desarrollado varias herramientas, sobre todo orientadas a obtener información valiosa sobre la evolución de los genomas de una especie. Entre ellas, las de uso más común son, en el caso de los procariontes Roary (Page, Cummins et al. 2015), Panseq (Laing, Buchanan et al. 2010), PGAweb (Chen, Zhang et al. 2018) y PanX (Ding, Baumdicker et al. 2018). En los eucariontes, aunque el uso del concepto de pan-genoma apenas comienza, existe la herramienta Pangloss (McCarthy and Fitzpatrick 2019).

Entre las posibles aplicaciones del análisis pan-genómico se encuentra el estudio de los factores de virulencia asociados a entidades patogénicas, o el diseño racional de vacunas. En el primer caso, se buscan los genes que correlacionan con la capacidad de los microorganismos de causar enfermedades, y entre otros ejemplos de su uso se pueden citar los trabajos de Doumith y colaboradores (Doumith, Cazalet et al. 2004) y de Schoen y colaboradores (Schoen, Blom et al. 2008). En el otro caso, la idea es encontrar proteínas comunes entre los representantes patogénicos de una especie, cuya poca variabilidad y su localización celular permitan su uso como antígenos en preparados vacunales.

En relación al segundo punto, se debe notar que la mayoría de las vacunas humanas contienen patógenos atenuados o muertos y fueron desarrolladas empíricamente, como la vacuna contra la fiebre amarilla (Rueckert and Guzmán 2012). Preocupaciones sobre la inocuidad asociada con preparados de vacunas basados en patógenos enteros ha llevado a que las nuevas vacunas se basen en un número restringido de componentes individuales del patógeno específico, que son capaces de conferir inmunidad protectora. Obviamente, las posibilidades de encontrar empíricamente componentes efectivos para estas vacunas son bajas. Las vacunas diseñadas racionalmente están compuestas de antígenos, sistemas de administración y a menudo adyuvantes que provocan respuestas inmunitarias predecibles contra epítopos específicos para proteger contra un patógeno en particular (ídem).

La selección del antígeno óptimo representa la piedra angular en el diseño de la vacuna (ibídem). Sobre todo, se buscan proteínas ubicadas en la membrana celular más externa y con segmentos expuestos que pudieran servir como potenciales antígenos (Perry and Ho 2013). Con el advenimiento de la genómica, el proceso tradicional de selección de antígenos candidatos uno por uno ha sido reemplazado por enfoques de vacunación inversa. Es decir, el potencial de codificación del genoma de un patógeno se explota mediante la selección *in silico*, los análisis de alto rendimiento y las tecnologías de elaboración de perfiles (por ejemplo, la genómica y la proteómica) para definir antígenos prometedores en relación con los genes expresados *in vivo* y la variación clonal (Rueckert and Guzmán 2012). Es en este punto en el cual los análisis pan-genómicos pueden brindar herramientas que faciliten la selección de un antígeno óptimo.

En este trabajo se presenta CGWork (del inglés Common Genome Workbench), una herramienta para el estudio de genómica comparativa con aplicaciones en el diseño racional de vacunas. El propósito de su desarrollo fue brindar a los investigadores una herramienta de fácil manejo para la búsqueda, en un conjunto de genomas de una especie bacteriana de interés, proteínas con un comportamiento evolutivo deseable (dígase, presentes en todos los organismos

y con baja variabilidad en sus determinantes antigénicos) para su uso como antígenos en preparados vacunales con el carácter más universal posible (que abarque la mayor cantidad de variantes del patógeno).

Materiales y métodos

CGWork se basa en la detección de conglomerados de genes ortólogos dado un conjunto de genomas, lo cual conlleva a un proceso de agrupamiento basado en una función de similitud entre secuencias biológicas. Este proceso se ramifica en dos aristas: la primera, optimizada para, con la mayor rapidez posible presentar resultados para un estudio preliminar del problema y, la segunda, optimizada para un estudio exhaustivo de los datos teniendo en cuenta su importancia biológica, mostrando con mayor precisión los resultados.

También presenta links de interés para seguir con una investigación.

Situación problema

Dado un conjunto de *N* genomas S = $\{G_1, \dots, G_N\}$, pertenecientes a una misma clasificación taxonómica (con G_i= $\{g_{i,1}, \dots, g_{i,N_i}\}$ como conjunto de genes del genoma i) se requiere realizar un análisis para extraer:

1. Los genes característicos de dicha clasificación (genoma núcleo o común, $\cap_1^N G_i$).
2. Los genes únicos de un subconjunto S_n de S, $\cap_{G_i \in S_n} G_i$ (incluye la búsqueda de los genes únicos de una especie *i*).
3. Los genes ortólogos con un gen *j* contenido en la especie *i*.

Algoritmo de agrupamiento

Todo proceso de agrupamiento requiere al menos dos componentes: una medida de la distancia (o similitud) entre los objetos a agrupar y un procedimiento de generación de grupos.

En el presente trabajo se emplearon dos enfoques diferentes para las distancias, uno que permite un cálculo rápido de las similitudes entre dos secuencias utilizando descomposición de palabras y uno basado en alineamiento para una búsqueda exhaustiva.

El primero es una aproximación al método USEARCH (Edgar 2010) y consiste en realizar primero la descomposición en k-tuplas de cada secuencia proteica a comparar, para luego comparar la biblioteca de k-tuplas de dos secuencias, hallando el tamaño de su intersección. La longitud de palabra *k* se toma como 5 por defecto. La similitud entre dos secuencias se mide como el porciento que representa la intersección con respecto al tamaño de la biblioteca de menor dimensión:

$Sim_{a,b} = \frac{|B_a \cap B_b|}{|B_b|}$ donde *a* y *b* son secuencias, B_a y B_b representan los conjuntos de k-tuplas, $|...|$ se usa como símbolo de dimensión y $|B_b|$ es la dimensión de la biblioteca de menor tamaño, derivada de la secuencia más corta *b*.

En el caso de las secuencias biológicas, una de las formas más comunes de establecer distancias entre ellas es a partir del puntaje de alineamiento. El alineamiento consiste en establecer una comparación entre dos cadenas de texto que representan a dos secuencias biológicas (sucesión de nucleótidos en el caso del ADN y ARN y de aminoácidos en el caso de las proteínas) de tal manera que se hagan coincidir las posiciones equivalentes, o en caso de no existir una posición equivalente se coloca un espacio vacío. El puntaje de este alineamiento es una función que le asigna un valor a cada par de posiciones que se hagan coincidir (sustituciones) más una función de penalización que depende de los espacios vacíos introducidos. El valor de los puntajes de sustitución normalmente se extrae de matrices llamadas matrices de sustitución. La matriz más sencilla de este tipo es la matriz identidad, la cual evaluaría con el valor 1 las coincidencias, y 0 las no coincidencias. Otras matrices más complejas se basan en enfoques frecuentistas o en fundamentos químico-físicos, y celdas portan información sobre la relevancia evolutiva y químico-biológica de la sustitución de una letra por otra (un nucleótido por otro en el caso de ácidos nucleicos, y de un aminoácido por otro en el caso de proteínas). En el Material Suplementario-Figura 1. se muestra un ejemplo de las matrices BLOSUM (Henikoff and Henikoff 1992).

Partiendo del puntaje de alineamiento, se construye una medida de similitud, representada por:

$Sim^M_{a,b} = \frac{P^M_{a,b}}{P^M_{a,a}}$, donde *a*, *b* son dos secuencias con longitud(a)> longitud(b), y $P^M_{a,b}$ se refiere al puntaje de alineamiento entre a y b usando la matriz M de sustitución. $P^M_{a,a}$ es el resultado de evaluar el alineamiento de una secuencia contra ella misma y representa el máximo puntaje a obtener en un alineamiento contra la secuencia *a*. Para

la obtención de los alineamientos, se usa el algoritmo de programación dinámica de Needleman-Wunsch (Needleman and Wunsch 1970) para la obtención de alineamientos globales.

En cuanto al procedimiento de agrupación, a diferencia de otros métodos basados para la detección de ortólogos (Remm, Storm et al. 2001; Galpert, Fernández et al. 2018) en los que se sigue un enfoque de todos contra todos (lo cual obliga al uso de capacidades de cómputo de alto rendimiento) en el presente trabajo se propone una variante de agrupamiento progresivo que guarda similitudes con el método de alineamiento múltiple CLUSTAL (Larkin, Blackshields et al. 2007).

El procedimiento se describe de la siguiente manera:

1. Se ordenan las secuencias en orden decreciente de longitud.
2. Se selecciona la secuencia de mayor longitud como semilla del primer conglomerado y se ejecuta el paso 3.
3. Dada la semilla de un conglomerado (*k*):
    i. Se realiza la comparación (la determinación de la intersección en las bibliotecas de k-tuplas en el caso rápido, o el alineamiento global usando Needleman-Wunsch en el caso exhaustivo) de esta secuencia con las restantes secuencias con *j*>*i*, donde *i* y *j* representan los índices de la secuencia semilla y la secuencia a comparar, respectivamente, en el ordenamiento inicial y se calcula la similitud ($Sim_{a,b}$ o $Sim^{M}_{a,b}$ según sea el caso) para cada comparación
    ii. Todas las secuencias cuyo valor de ($Sim_{a,b}$ o $Sim^{M}_{a,b}$) esté por encima de un valor de corte C_{sim} predefinido, se adicionan al conglomerado *k*.
4. Se busca la secuencia de mayor índice que aún no esté incluida en ningún conglomerado:
    i. Si esta condición es nula se finaliza
    ii. Si existe alguna secuencia que cumpla la condición, se incrementa *k*, se hace a esta secuencia la semilla de este nuevo conglomerado y se va al paso 3.

Descripción de la herramienta

Como se describió en el planteamiento del problema, en el programa, una vez obtenido un agrupamiento, pueden realizarse diferentes operaciones como buscar genes que son característicos de un genoma o de un conjunto de ellos, o buscar los genomas donde aparecen genes homólogos de un gen en particular.

Mediante la utilización de la herramienta se posibilita este proceso con la búsqueda de agrupamientos que contengan a todos los genomas de un subconjunto S_n. Se encuentra a dicho conjunto de genes mediante $\cup_{C_k \cap G_i \neq \emptyset, \forall G_i \in S_n} C_k$, es decir, la unión de todos los clústeres C_k que contengan intersección de genes con todos los genomas de S_n. Un caso particular sería cuando se quiere buscar el genoma núcleo o genoma común a todo el taxón ($S_n = S$), representado por $\cap_1^N G_i$. Para ello se busca $\cup_{C_k \cap G_i \neq \emptyset, \forall G_i \in S} C_k$. En muchas ocasiones, el problema biológico particular implica localizar los genes que pertenecen a un subconjunto de genomas, y no están presentes en el resto. En tal situación, la búsqueda se hace como $\cup_{C_k \cap G_i \neq \emptyset \wedge C_k \cap G_j = \emptyset, \forall G_i \in S_n, G_j \notin S_n} C_k$, esta búsqueda se tratará en lo adelante como búsqueda estricta, mientras la anterior se trata como búsqueda abierta. Cuando solo se quiere buscar los genes característicos de un único genoma *i* ($S_n = \{G_i\}$), bastaría solo con realizar una búsqueda de clústeres que solo contengan a dicho genoma, por lo tanto, es un caso particular de una búsqueda estricta.
Cuando se quiera buscar los genomas donde aparecen genes homólogos de un gen en particular, basta localizar el clúster que contiene al gen, $C_k \ni g_{i,j}$ y determinar los genomas que contienen algún gen en el clúster ($G_i : C_k \cap G_i \neq \emptyset$).

Para facilitar a los especialistas en ciencias de la vida la realización de estas operaciones, se programó una herramienta utilizando el lenguaje de programación Java 8.0.0, usando como herramienta de desarrollo el Apache NetBeans IDE 8.2. En el Material Suplementario-Figura 2. se puede ver la pantalla principal de la aplicación, y su descripción.

Caso de estudio

La meningitis meningocócica es una infección bacteriana grave de las membranas que rodean el cerebro y la médula espinal. Puede causar importantes daños cerebrales y es mortal en el 50% de los casos no tratados. Aunque existen vacunas efectivas contra la mayoría de los serogrupos de este patógeno, en el caso del serogrupo B, las vacunas basadas en vesículas de membrana externa (OMV), que son generalmente cepa-específicas, continúan siendo las de mejores resultados.

Cuba cuenta con una vacuna de este tipo patentada en el 1991 que sirvió para contener la epidemia originada por la cepa CU385 a inicios de los 90 del pasado siglo y que forma parte del programa de vacunación nacional. Dado que se

observaron respuestas de reactividad cruzada contra otras cepas del serogrupo B inducidas por este preparado vacunal, el interés por este posibilita además que se continúe su comercialización a nivel internacional. Recientemente como parte de un estudio europeo el genoma de esta cepa se secuenció y publicó en la base de datos de la NCBI por segmentos. Sin embargo, no se ha realizado una caracterización genómica profunda que permita su comparación con otras cepas utilizadas en la preparación de vacunas de OMV a nivel internacional. La información obtenida por esta investigación respaldará el expediente de la vacuna cubana y permitirá la identificación de genes de valor inmunogénico y microbiológico.

Por otra parte, el CIGB sistemáticamente ha ido caracterizando mediante herramientas de la Proteómica la composición proteica del proteoliposoma que constituye el ingrediente farmacéutico activo de la vacuna VAMENGOC-BC (Betancourt, Gil et al. 2005; Uli, Castellanos- Serra et al. 2006; Ramos, Gutierrez et al. 2008; Gil, Betancourt et al. 2009; Masforrol, Gil et al. 2017). Todas las búsquedas se realizaron en su momento contra una base de datos de secuencias de una cepa de referencia de **Neisseria meningitidis** que se había secuenciado (MC58), sin embargo el disponer en estos momentos de los datos primarios de Proteómica y la secuencia del genoma de la cepa CU385, además de varios genomas del serogrupo B de **Neisseria meningitidis**, nos permite realizar las búsquedas apropiadas y confirmar las anotaciones realizadas en el genoma así como la identificación de proteínas de membrana presentes en la OMV de la vacuna cubana que justifiquen la reactividad cruzada. Basado es este problema y los datos viables en el momento se realizó una investigación utilizando la herramienta expuesta.

Resultados y discusión

El uso de la herramienta comienza por la selección de las opciones de agrupamiento (Material Suplementario-Figura 3.), que incluye los archivos de entrada, el tipo de agrupamiento y los parámetros de agrupamiento en dependencia del tipo. Una vez que se seleccionan las opciones y se presiona el botón "Submit options" automáticamente se ejecuta el agrupamiento.

Una vez realizado el agrupamiento, el usuario puede realizar la selección de los genomas con los que desea trabajar, así como el tipo de búsqueda, incluyendo el uso de operadores lógicos (Material Suplementario-Figura 4). Las operaciones seleccionadas aparecen en la barra de operaciones, y se ejecuta la búsqueda presionando el botón "browse". También puede hacer la búsqueda de los genomas donde se incluye un determinado gen mediante la barra da búsqueda.

Los resultados se muestran en texto simple, imprimiendo el nombre del clúster y los identificadores del gen y genoma correspondiente. Si se quieren realizar operaciones posteriores como la anotación funcional usando algunos de los programas disponibles online, se necesitan estas secuencias en formato FASTA, así que la herramienta permite generar la salida en este formato por cada clúster, lo cual se hace automáticamente al seleccionar este formato de salida. Si se quiere estudiar la variabilidad de las proteínas dentro de un clúster, se necesita realizar el alineamiento múltiple de las secuencias en el mismo. Cuando el usuario selecciona la opción de salida alineamiento múltiple ("multiple alignment") se hace una llamada a MUSCLE (Edgar 2004) y se obtiene el alineamiento múltiple dentro de cada clúster en la salida.

Para el diseño racional de vacunas, los criterios de filtrado de las proteínas se realizan en el orden: Conjunto de organismos diana (contra los que va dirigida la vacuna)/Proteínas comunes a todos/Proteínas de la membrana más externa/Segmentos de la proteína expuestos en la superficie. Para realizar este filtrado, en el presente trabajo se propone el flujo: Filtrado de genomas con CGWork (Selección de los genomas de interés) → Búsqueda del genoma núcleo con CGWork → Predicción de proteínas de membrana más externa PSORTb3.0 (Yu, Wagner et al. 2010) / Predicción de funciones y dominios usando BLAST (Altschul, Gish et al. 1990) → Predicción de lazos externos con Phobius (Käll, Krogh et al. 2007) → Análisis de variabilidad en CGWork de los segmentos predichos como lazos. En el estado actual de la herramienta, las predicciones con PSORTb3.0, BLAST y Phobius se realizan en línea, usando los servidores disponibles en Internet, por lo que se proveen los vínculos a estas herramientas.

Para el caso de estudio de la reactividad cruzada de la vacuna cubana contra la Meningitis Meningocóccica se utilizó la herramienta CGWork para una investigación preliminar del problema, donde, una vez seleccionado el conjunto de genomas disponibles del serogrupo B de **Neisseria meningitidis** se realizó el filtrado siguiendo dos enfoques: uno de ellos buscaba encontrar genes de membrana externa en el genoma núcleo cuya variabilidad en los segmentos expuestos justificara las diferentes respuestas a VAMENGOC-BC, y el otro buscaba genes de membrana externa sólo observables en el subconjunto de cepas que muestran reactividad cruzada a VAMENGOC-BC. Los resultados de este estudio inicial han brindado indicios adicionales a los investigadores del CIGB sobre las posibles causas de la reactividad cruzada a VAMENGOC-BC.

El estudio se realizó en una laptop HP Radeon A-300 a 1300 Hz, con 1Gb de RAM. Con un total de 18 genomas, con 1872 proteínas por genoma como promedio, utilizando el proceso de agrupamiento exhaustivo el procesamiento demoró 6 horas.Nótese que bajo una capacidad de procesamiento baja e inferior a las utilizadas normalmente con

respecto a las especificaciones de los softwares homólogos a este, muestra un incremento de velocidad de procesamiento.

Conclusiones

La herramienta CGWork permite la selección de proteínas en un conjunto de genomas que cumplan con un criterio determinado y que puede especificarse mediante operaciones lógicas. Basa su funcionamiento en un proceso de agrupamiento progresivo con dos medidas de similitud, una que permite el cálculo rápido con baja sensibilidad, basada en descomposición en k-tuplas, y otra que permite una búsqueda exhaustiva, basada en alineamientos, con una mayor sensibilidad, pero con mayor tiempo de procesamiento. La herramienta tiene potencialidades en el diseño racional de vacunas, y ya se está empleando en un proyecto cuyo fin es la explicación de la reactividad cruzada a la vacuna cubana contra meningitis meningocócica VAMENGOC-BC.

Agradecimientos

Los autores declaran un agradecimiento especial al profesor Msc. Mario Pupo del Departamento de Bioinformática, Universidad de Ciencias Informáticas por la tutoría en todo el proceso incluyendo la revisión del manuscrito y al investigador Luis Javier del Centro de Ingeniería Genética y Biotecnología por su especial atención y retroalimentación. Adicionalmente al colectivo de trabajadores y profesores del Departamento de Bioinformática, Universidad de Ciencias Informáticas.

Conflictos de intereses

Los autores de este artículo no declaran ningún conflicto de interés con respecto a lo expuesto en el mismo.

Contribución de los autores

1. Conceptualización: Jorge Alejandro Jiménez Garí, Fernando Figueredo Sánchez, Víctor Ventura Mena.
2. Curación de datos: Jorge Alejandro Jiménez Garí.
3. Análisis formal: Jorge Alejandro Jiménez Garí.
4. Investigación: Jorge Alejandro Jiménez Garí
5. Metodología: Fernando Figueredo Sánchez, Víctor Ventura Mena.
6. Software: Fernando Figueredo Sánchez, Víctor Ventura Mena.

7. Supervisión: Jorge Alejandro Jiménez Garí.
8. Validación: Fernando Figueredo Sánchez, Víctor Ventura Mena.
9. Visualización: Fernando Figueredo Sánchez, Víctor Ventura Mena.
10. Redacción – borrador original: Jorge Alejandro Jiménez Garí, Fernando Figueredo Sánchez, Víctor Ventura Mena.
11. Redacción – revisión y edición: Jorge Alejandro Jiménez Garí, Fernando Figueredo Sánchez, Víctor Ventura Mena.

Referencias

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Betancourt, L., J. Gil, et al. (2005). "SCAPE: a new tool for the Selective CApture of PEptides in protein identification." Journal of proteome research **4**(2): 491-496.

Chen, X., Y. Zhang, et al. (2018). "PGAweb: a web server for bacterial pan-genome analysis." Frontiers in microbiology **9**: 1910.

Ding, W., F. Baumdicker, et al. (2018). "panX: pan-genome analysis and exploration." Nucleic Acids Research **46**(1): e5-e5.

Dobrindt, U. and J. Hacker (2001). "Whole genome plasticity in pathogenic bacteria." Current Opinion in Microbiology **4**(5): 550-557.

Doumith, M., C. Cazalet, et al. (2004). "New Aspects Regarding Evolution and Virulence of Listeria monocytogenes Revealed by Comparative Genomics and DNA Arrays." Infection and Immunity **72**(2): 1072-1083.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-1797.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edwards, D. J. and K. E. Holt (2013). "Beginner's guide to comparative bacterial genome analysis using next-generation sequence data." Microbial informatics and experimentation **3**(1): 2.

Galpert, D., A. Fernández, et al. (2018). "Surveying alignment-free features for Ortholog detection in related yeast proteomes by using supervised big data classifiers." BMC Bioinformatics **19**(1).

Gil, J., L. H. Betancourt, et al. (2009). "Proteomic study via a non-gel based approach of meningococcal outer membrane vesicle vaccine obtained from strain CU385: a road map for discovering new antigens." Human vaccines **5**(5): 347-356.

Hardison, R. C. (2003). "Comparative Genomics." PLoS Biology **1**(2): e58.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences **89**(22): 10915-10919.

Käll, L., A. Krogh, et al. (2007). "Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server." Nucleic Acids Research **35**(suppl_2): W429-W432.

Koonin, E. V. and Y. I. Wolf (2012). "Evolution of microbes and viruses: a paradigm shift in evolutionary biology?" Frontiers in cellular and infection microbiology **2**: 119.

Laing, C., C. Buchanan, et al. (2010). "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions." BMC Bioinformatics **11**(1): 461.

Lapierre, P. and J. P. Gogarten (2009). "Estimating the size of the bacterial pan-genome." Trends in genetics **25**(3): 107-110.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Masforrol, Y., J. Gil, et al. (2017). "A deeper mining on the protein composition of VA-MENGOC-BC®: An OMV-based vaccine against N. meningitidis serogroup B and C." Human vaccines & immunotherapeutics **13**(11): 2548-2560.

McCarthy and Fitzpatrick (2019). "Pangloss: A Tool for Pan-Genome Analysis of Microbial Eukaryotes." Genes **10**(7): 521.

Miller, W., K. D. Makova, et al. (2004). "Comparative Genomics." Annual Review of Genomics and Human Genetics **5**(1): 15-56.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." Journal of molecular biology **48**(3): 443-453.

Page, A. J., C. A. Cummins, et al. (2015). "Roary: rapid large-scale prokaryote pan genome analysis." Bioinformatics **31**(22): 3691-3693.

Perry, A. J. and B. K. Ho (2013). "Inmembrane, a bioinformatic workflow for annotation of bacterial cell-surface proteomes." Source code for biology and medicine **8**(1): 9.

Ramos, Y., E. Gutierrez, et al. (2008). "Proteomics based on peptide fractionation by SDS-free PAGE." Journal of proteome research **7**(6): 2427-2434.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." Journal of molecular biology **314**(5): 1041-1052.

Rueckert, C. and C. A. Guzmán (2012). "Vaccines: from empirical development to rational design." PLoS pathogens **8**(11): e1003001.

Schoen, C., J. Blom, et al. (2008). "Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis." Proceedings of the National Academy of Sciences **105**(9): 3473-3478.

Tettelin, H., D. Riley, et al. (2008). "Comparative genomics: the bacterial pan-genome." Current Opinion in Microbiology **11**(5): 472-477.

Uli, L., L. Castellanos- Serra, et al. (2006). "Outer membrane vesicles of the VA- MENGOC- BC® vaccine against serogroup B of Neisseria meningitidis: Analysis of protein components by two- dimensional gel electrophoresis and mass spectrometry." Proteomics **6**(11): 3389-3399.

Yu, N. Y., J. R. Wagner, et al. (2010). "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes." Bioinformatics **26**(13): 1608-1615.