Vector 11 (2016) 5 - 10
ISSN 1909 - 7891

Procesamiento y visualización distribuida de relaciones funcionales de genes en ambientes ubicuos

Cesar Augusto Franco Estrada*[a]

[a] *Ingeniero en Computación. Universidad de Caldas. Manizales - Colombia*

Recibido: 23/06/16. Aprobado: 18/10/16

Resumen

En esta investigación se presenta el resultado de la validación de la eficiencia de un prototipo computacional, el cual podrá permitir a los investigadores en ciencias de la vida, facilitar la realización de algunas tareas que están relacionadas con el procesamiento de la información biológica y se deben realizar con frecuencia; con esto podrán lanzar procesos de consulta de relaciones funcionales de genes de forma remota haciendo uso de la computación distribuida y ubicua, utilizando dispositivos que son comunes en nuestro entorno como son los teléfonos inteligentes y las tabletas con los cuales se puede consultar los resultados adquiridos de esos procesos expresados en representaciones graficas; en correspondencia se evidencia la disminución de tiempos en la ejecución de estas complejas tareas y se mejora la capacidad de interpretación de los resultados finales del pipeline bioinformático.
Se ejecutaron diferentes pruebas con el fin de determinar la eficiencia del procesamiento de un mismo trabajo realizado por un dispositivo móvil (proceso no distribuido) contra varios dispositivos móviles (proceso distribuido), para ello se tomaron varias medidas usando variables como consumo de batería y tiempo de procesamiento.

Palabras clave: Bioinformática, Computación Distribuida, Computación Ubicua, Visualización.

Distributed processing and visualization of functional relations of genes in ubiquitous environments

Abstract

This research presents the results to validate the efficiency of a computer prototype, which may allow researchers in life sciences, to facilitate performance of some tasks that are related to the processing of biological information and should be performed frequently; with this, they will be able to launch consultation processes of functional relationships of genes remotely using distributed and ubiquitous computing, integrating devices that are common in our environment such as smartphones and tablets with which they can view the results acquired from these processes, expressed in graphical representations; correspondingly, the decreasing execution time of these complex tasks is evident, and the interpretation capacity to the final results for the bioinformatics pipeline is improved. Different tests were performed in order to determine the efficiency of the processing of a same work performed by a mobile device (undistributed process) against multiple mobile devices (distributed process), for this process, some measures were obtained using variables such as battery consumption and processing time.

Key words: Bioinformatics, Distributed Computing, Ubiquitous Computing, Visualization.

1. Introducción

En los últimos años el área de la genética ha evolucionado al punto de generar grandes volúmenes de datos heterogéneos de las especies vivientes que puede ayudar a comprender mejor su comportamiento, sin embargo, el proceso de análisis es muy dispendioso para tratarse manualmente in-vitro, ya que debe procesarse con técnicas intensivas y algoritmos que consumen elevados recursos. Por tal motivo se han involucrado diferentes áreas de las ciencias naturales tales como: la matemática, la física, la química y la biología, lo que ha permitido avanzar en la obtención de resultados y ha ayudado a agilizar y mejorar estos procesos disminuyendo el tiempo requerido para la obtención de los mismos. Estas áreas han explorado herramientas como la tecnología la cual ha avanzado rápidamente durante los últimos años facilitando así el desarrollo de diferentes procesos, y gracias a estos avances y a la articulación de las diferentes ciencias se ha logrado crear la bioinformática.

* *E-mail: cesar.franco@ucaldas.edu.co (C. Franco)*
orcid.org/0000-0003-4863-1728

Cómo citar este artículo:
Franco Estrada C.A. (2016). Procesamiento y visualización distribuida de relaciones funcionales de genes en ambientes ubicuos. *Revista Vector*, 11: 5-10.

En el campo de la bioinformática se han adelantado una serie de investigaciones que han generado conocimiento acerca de la composición genética de organismos a nivel molecular y se han almacenado en bases de datos ontológicas como *Gene Ontology* (Ontology, 2015), que es uno de los recursos principales de la información biológica, ya que proporciona una definición específica sobre las funciones de la proteína (Guzzi *et al.*, 2014). La bioinformática ha sido de gran importancia en el avance científico, ya que ha permitido realizar análisis y procesamiento de datos biológicos a través de herramientas y algoritmos computacionales. (Arango L., 2014).

Con el gran auge de la bioinformática en los últimos años y una nueva era Post-PC (Murtagh, 2014), se hace necesario investigar nuevas formas para realizar procesamiento distribuido genómico y visualizar los datos a través de dispositivos tales como: teléfonos inteligentes y tabletas, los cuales se están masificando a una gran velocidad y afectando directamente a las ventas de computadoras de escritorio y portátiles. En la Figura 1 se puede observar la disminución porcentual en las ventas de computadoras personales comparando los años 2011–2012 (Gartner, 2013), 2012–2013 (Gartner, 2014) y 2013–2014 (Gartner, 2015).

Esto no quiere decir que las personas hayan dejado de adquirir tecnología, al contrario se ha visto un aumento significativo en la compra de dispositivos móviles como *smartphones*, *tablets*, muchos de los cuales suplen ciertas funcionalidades de un PC, además al ser fácilmente transportables también se utilizan como medios de comunicación e interconectividad en cualquier sitio. Estos equipos se pueden manejar de forma independiente y al haber tantos en el mercado se pueden aprovechar para realizar tareas de forma distribuida.

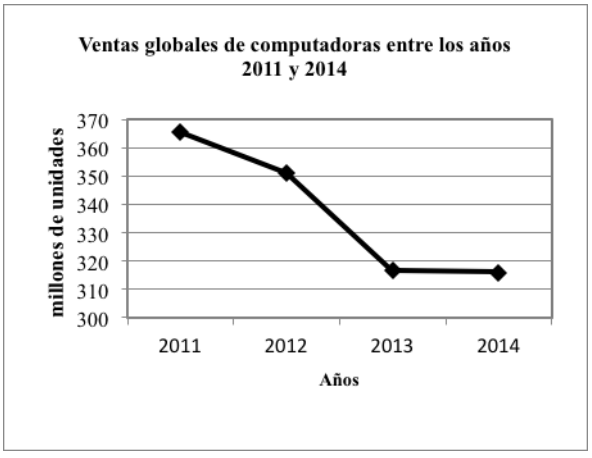

Figura 1. *Ventas globales de computadoras entre los años 2011 y 2014.*
Fuente: *Gartner (2013, 2014, 2015).*

En muchos contextos se distribuyen los datos necesarios para realizar las tareas de procesamiento, por lo que una solución distribuida para procesar los datos parece natural y podría ser más eficiente. Sin embargo, la computación distribuida también trae una serie de desafíos, por ejemplo, se necesita algún mecanismo para hacer un seguimiento de los equipos que son relevantes para un determinado cálculo, sobre todo sí este sistema de computadoras cambia a lo largo del tiempo, por ejemplo, diferentes computadoras pueden almacenar datos que pueden ser necesarios en diferentes momentos durante el cálculo, o algunos equipos pueden llegar a ser más relevante / irrelevante debido los problemas de balance de carga (Bobed *et al.*, 2013).

En la actualidad se puede identificar que las personas buscan productos que les faciliten el uso de su tiempo y rapidez al momento de realizar diferentes tareas, por lo que requieren el uso de equipos que sean fácilmente transportables. De acuerdo a los avances tecnológicos, a las tendencias de compra de los usuarios y a las mismas necesidades y usos que las personas les están dando a los dispositivos móviles se hace necesario empezar a crear métodos que faciliten el desarrollo de procesos, en este caso el análisis de información genómica, transcriptómica entre otras. El uso de los diferentes dispositivos móviles puede contribuir a agilizar el análisis y la obtención de resultados en el campo de la biología, ya que los investigadores podrían ejecutar diferentes tareas de forma remota y disminuirían los tiempos de respuesta, optimizando así el procesamiento y análisis de datos.

2. Estado del arte

Si bien los resultados de investigación en el uso intensivo y distribuido de la computación ubicua en ambientes bioinformáticos es apenas incipiente, al momento de realizar la exploración de los trabajos de investigación se encontraron indicios de proyectos que muestran un interés por diseñar, implementar y evaluar prototipos que permitan realizar procesos distribuidos usando la computación ubicua, esto se debe a la necesidad de aprovechar la gran cantidad y disponibilidad de herramientas y de dispositivos móviles; es de anotar que el número de usuarios de estos equipos superó al número de usuarios de computadoras de escritorio (Murtagh, 2014). Trabajos como: *Mobile MapReduce: Minimizing Response Time of Computing Intensive Mobile Applications* (Hassan y Chen, 2011) que investiga cómo minimizar efectivamente el tiempo de respuesta de los usuarios de aplicaciones móviles, *Data Clustering on a Network of Mobile*

Smartphones (Kalogeraki *et al.*, 2011) que presenta un sistema para la agrupación de datos en una red de teléfonos inteligentes móviles, *A Review on Distributed Application Processing Frameworks in Smart Mobile Devices for Mobile Cloud Computing* (Shiraz *et al.*, 2012) que destaca los problemas y desafíos con los *Frameworks* existentes para el procesamiento de aplicaciones distribuidas en el desarrollo, implementación y ejecución de aplicaciones móviles computacionalmente intensivas dentro del dominio de computación en la nube móvil, son evidencia del aprovechamiento de las bondades del uso de los dispositivos.

También se puede identificar en la web que ya se ha iniciado el desarrollo de aplicaciones y plataformas para el realizar procesamiento distribuido o búsquedas y consultas relacionadas con información biológica usando dispositivos móviles. A continuación se encontrará una descripción de las funcionalidades de cada una, aunque muchas de estas ya no están disponibles.

3. Materiales y métodos

Para la construcción del prototipo se usó como proceso de desarrollo de software el modelo "En Cascada" (Waterfall), ya que este modelo es apropiado para soluciones en los que los requerimientos establecidos son cerrados y no van a cambiar durante las fases en la que se divide el proyecto, muy aconsejable para proyectos altamente controlados y predecibles. Durante este proceso se implementaron 3 módulos los cuales se describen a continuación:

3.1 Implementación de un módulo de administración:

Para el proceso de administración de trabajos, tareas y dispositivos se desarrolló un módulo que permite asignar a los dispositivos móviles las secuencias de proteínas que se deben enviar para su posterior procesamiento. Este administrador de tareas está basado en un conjunto de Servicios Web que proporcionarán los mecanismos de comunicación estándares para interactuar entre los dispositivos móviles facilitando la interoperabilidad y extensibilidad entre estos.

3.2 Implementación de un módulo para el procesamiento distribuido:

Para el proceso de búsqueda de secuencias en bases de datos de firmas de proteínas se desarrolló un módulo que permite a cada dispositivo móvil realizar esta tarea de forma independiente, permitiendo a cada dispositivo conectarse al servicio externo de búsqueda InterProScan y obtener resultados.

3.3 Implementación de un módulo para la visualización de los resultados:

Para la representación de los resultados obtenidos se desarrolló un módulo para la visualización que permite identificar la relaciones encontradas usando gráficos de tipo "*Radial Tree*", "*Collapsible Tree*" y "*Nodes*" para ello se utilizó D3.js (Bostock, 2015).

3.4 Servicio externo de apoyo para el análisis de proteínas:

La base de datos InterPro integra modelos predictivos o firmas de múltiples y diversas fuentes de repositorios: *Pfam* (Punta *et al.*, 2010), *PRINTS* (Attwood *et al.*, 2003), *PROSITE* (Sigrist *et al.*, 2010), *SMART* (Letunic *et al.*, 2009), *ProDom* (Bru *et al.*, 2005), *PIRSF* (Nikolskaya *et al.*, 2006), *SUPERFAMILY* (de Lima Morais *et al.*, 2011), *PHANTER* (Mi *et al.*, 2010), *CATH-Gene3D* (Lees *et al.*, 2010), *TIGRFAMs* (Selengut *et al.*, 2007) y *HAMAP* (Lima *et al.*, 2009). Cada fuente tiene su propio enfoque biológico distinto y/o metodología de la producción de la firma. El objetivo de *InterPro* es combinar sus fuerzas individuales para proporcionar un único recurso a través del cual los científicos pueden acceder a la información completa acerca de las familias de proteínas, dominios y sitios funcionales (Hunter *et al.*, 2012).

4. Resultados y discusión

4.1 Computacional

Con el desarrollo de la investigación se logró obtener un prototipo computacional que permite la distribución y visualización de las relaciones funcionales de genes en dispositivos móviles. En la Figura 2 se puede observar el diagrama que muestra una visión general del comportamiento del prototipo.

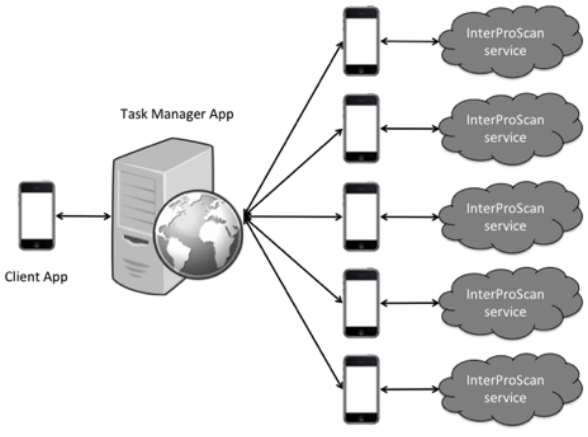

Figura 2. *Visión general del comportamiento del prototipo.*

El prototipo inicia preparando los datos que se requieren para la creación del flujo de trabajo y se envía al administrador de tareas para que sea registrado y esté disponible para su posterior procesamiento, que será realizado por los dispositivos móviles registrados.

Después de haber creado el flujo de trabajo cada uno de los dispositivos móviles registrados solicita al administrador de tareas una actividad disponible para ser procesada; el administrador verifica la identidad del dispositivo y asigna una tarea para que el dispositivo pueda obtener la secuencia de proteínas.

Cuando el dispositivo móvil obtiene la secuencia de proteínas correspondiente a la tarea asignada establece una comunicación con el servicio *InterProScan* y a través de un servicio Web *RESTful* envía la secuencia para su procesamiento. El procesamiento de este servicio es en segundo plano por eso es necesario utilizar otro de los servicios Web que tiene a disposición esta plataforma para consultar el estado del proceso hasta que este haya finalizado.

Cuando el servicio *InterProScan* reporta que el proceso ha concluido se invoca nuevamente otro servicio Web para obtener la respuesta a la consulta realizada. Esta respuesta es enviada al administrador de tareas para que sea almacenada en el servidor y la tarea sea marcada como realizada. Posteriormente el dispositivo vuelve a solicitar otra tarea y se repiten los pasos anteriores hasta realizar todas las tareas y finalizar el flujo de trabajo.

Después de registrar cada tarea se genera un archivo con la información de las relaciones funcionales a través del cual se puede obtener un gráfico donde se visualice cual es la respuesta que se generó. Dichos datos tienen relación con familias, dominios y sitios funcionales de proteínas almacenados en la base datos utilizada *InterPro.* Los tipos de gráficos seleccionados por su estructura en forma de árbol o nodos muestran las relaciones de una forma que facilitan la interpretación por parte de los expertos en el área de la bioinformática.

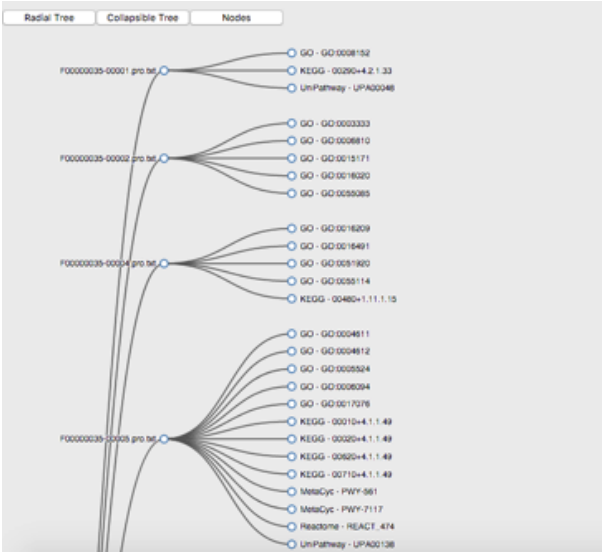

Figura 3. *Presentación de resultados usando un gráfico tipo "Collapsible Tree"*

4.2 Pruebas

Con el prototipo se realizaron diferentes pruebas tanto de un modelo distribuido, como de un modelo no distribuido, con el fin de determinar la eficiencia de trabajo de un equipo comparado con el uso de varios dispositivos para el mismo análisis. Se contabilizó el tiempo total requerido para las 297 tareas que se recibieron y la comparación de la eficiencia del modelo se visualiza en la Figura 4.

Se tomaron variables como tiempo total de procesamiento y consumo de batería para la validación de los resultados de las pruebas que se realizaron con los diferentes dispositivos utilizados. Éstas se consideraron relevantes porque permiten identificar si la ejecución de las pruebas es eficiente o no frente al tiempo y su impacto directo en el consumo de batería, lo que puede afectar la autonomía del equipo, debido a que este es uno de los elementos esenciales en los dispositivos móviles actuales.

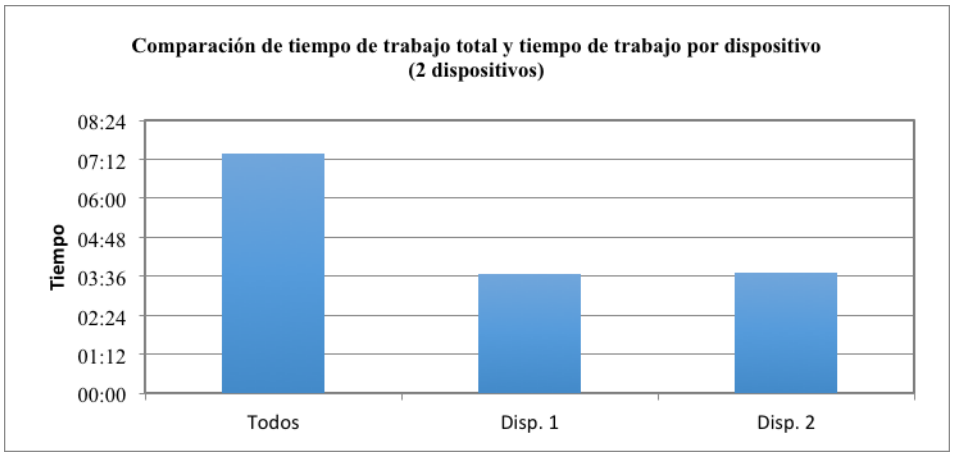

Figura 4. *Comparación de tiempos de trabajo.*

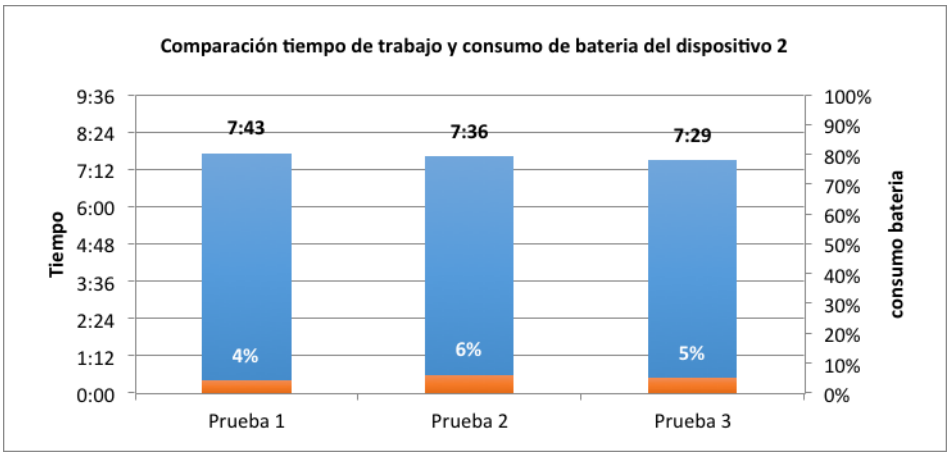

Figura 5. *Consumo de tiempo y batería del dispositivo Moto G X1032.*

4.3 Biológicos

Este proyecto se desarrolló utilizando un conjunto de datos donde se encuentran almacenadas las secuencias de ADN del organismo *Ganoderma Australe*, proporcionado por el grupo de investigación GITIR de la Universidad de Caldas. Este archivo ha sido anotado y ensamblado con la herramienta BLAST2GO y se obtuvieron una cantidad de 10.258 *contigs* para analizar. Para la prueba de la funcionalidad de la aplicación se tomó como prueba este organismo, pero puede ser usada para el procesamiento de otros de forma general.

Para el desarrollo se requiere preparar los datos (secuencias de ADN) a través de algoritmos computacionales con los cuales realiza un proceso de división, filtrado, transcripción, traducción y recorte requeridos; después de dicho proceso se logró obtener 297 secuencias de proteínas adecuadas para ser enviadas y analizadas por el servicio externo *InterProScan*.

5. Conclusiones

Para implementar un prototipo computacional de distribución de las relaciones funcionales de genes en dispositivos móviles, se desarrolló un software compuesto por dos módulos distribuidos así: el primero módulo es un administrador de tareas implementado como un aplicación Web que permite llevar un control de los trabajos, tareas y dispositivos que las realizan; el segundo módulo es un cliente implementado como una aplicación para dispositivos móviles que permite enviar secuencias de ADN, recibir secuencias de proteínas y realizar búsquedas estableciendo una comunicación con el servicio externo *InterProScan* para registrar el resultado.

Para analizar y diseñar un módulo de visualización de relaciones funcionales de genes para ambientes móviles derivado de un procesamiento distribuido, se desarrolló un módulo de visualización que permite leer un archivo generado por el prototipo con los resultados de las relaciones encontradas después

del procesamiento distribuido. Con dicho archivo se pueden presentar una serie de gráficos donde se visualizan las respuestas que se obtuvieron del procesamiento de cada secuencia de proteínas y allí se puede identificar las coincidencias encontradas en las diferentes bases de datos que se encuentran asociadas al servicio *InterProScan*.

Por otra parte, en la revisión sistemática del estado del arte, se ha encontrado precariedad e inmadurez en las librerías y estándares de descomposición paralela en ambientes de desarrollo para dispositivos ubicuos, en ese sentido, fue necesario implementar funciones y métodos propios que permitieran ejecutar este tipo de tareas, no obstante, en la medida que estos artefactos computacionales se extiendan, se podrían incorporar al prototipo. También, se encontraron algunos trabajos de investigación que dan un inicio a la construcción de *frameworks*, que podrían ser útiles para la construcción de aplicaciones que requieran computación distribuida en ambientes ubicuos, pero estos desarrollos tienen limitaciones que no permiten ser utilizados para solucionar cualquier tipo de problema.

Agradecimientos

El presente trabajo de investigación fue realizado bajo la supervisión del grupo de investigación GITIR, Ph.D. Gustavo A. Isaza E. y Ph.D. Luis F. Castillo O.

Referencias

Arango L., J. (2014). *Representación Ontólogica de Genes Basados en Linked Open Data.*

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., . . . Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS.

Bobed, C., Ilarri, S., & Mena, E. (2013). Distributed Mobile Computing: Development of Distributed Applications Using Mobile Agents.

Bostock, M. (2015). Introduction to Data-Driven Documents.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D.

de Lima Morais, D. A., Fang, H., Rackham, O. J. L., Wilson, D., Pethica, R., Chothia, C., & Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene ontology method.

Gartner. (2013). Gartner Says Declining Worldwide PC Shipments in Fourth Quarter of 2012 Signal Structural Shift of PC Market.

Gartner. (2014). Gartner Says Worldwide PC Shipments Declined 6.9 Percent in Fourth Quarter of 2013.

Gartner. (2015). Gartner Says Worldwide PC Shipments Grew 1 Percent in Fourth Quarter of 2014.

Guzzi, P. H., Milano, M., & Cannataro, M. (2014). Mining Association Rules from Gene Ontology and Protein Networks: Promises and Challenges. *Procedia Computer Science*.

Hassan, M. A., & Chen, S. (2011). Mobile MapReduce: Minimizing Response Time of Computing Intensive Mobile Applications. *Mobile Computing, Applications, and Services*.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., . . . Yong Siew-Yit. (2012). InterPro in 2011: new developments in the family and domain prediction database.

Kalogeraki, V., Gunopulos, D., Mielikinen, T., Tuulos, V. H., Foley, S., & Yu, C. (2011). Data Clustering on a Network of Mobile Smartphones. *Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on*.

Lees, J., Yeats, C., Redfern, O., Clegg, A., & Orengo, C. (2010). Gene3D: merging structure and function for a Thousand genomes.

Letunic, I., Doerks, T., & Bork, P. (2009). SMART 6: recent updates and new developments.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., . . . Bairoch, A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., & Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium.

Murtagh, R. (2014). Mobile Now Exceeds PC: The Biggest Shift Since the Internet Began. *Search Engine Watch*. Retrieved from https://searchenginewatch.com/sew/opinion/2353616/mobile-now-exceeds-pc-the-biggest-shift-since-the-internet-began

Nikolskaya, A. N., Arighi, C. N., Huang, H., Barker, W. C., & Wu, C. H. (2006). PIRSF family classification system for protein functional and evolutionary analysis.

Ontology, T. G. (2015). Gene Ontology Documentation. Retrieved from http://www.geneontology.org/page/documentation

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., . . . Finn, R. D. (2010). The Pfam protein families database.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., . . . White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes.

Shiraz, M., Gani, A., Khokhar, R. H., & Buyya, R. (2012). A Review on Distributed Application Processing Frameworks in Smart Mobile Devices for Mobile Cloud Computing.

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation.