

Cloud Computing con Arcimboldo para una solución de estructuras cristalográficas

Objetivo del proyecto

La cristalografía macromolecular, que proporciona una visión de la estructura tridimensional de las macromoléculas biológicas con un altísimo grado de detalle y precisión, trabaja a partir de datos de difracción medidos en sincrotrón (acelerador de partículas donde se emplea un campo magnético para acelerar un haz de partículas con carga eléctrica y mantenerlas en una órbita cerrada).

Con Arcimboldo se pretende dar solución a la falta de herramientas informáticas y a la complejidad del acceso por parte de los usuarios a los centros de supercomputación, ya que el desarrollo de software establecido para biología estructural se amplía a la explotación remota de cálculo masivo en paralelo. De esta forma la comunidad científica tendrá a su alcance una solución global de software, middleware y hardware a un problema concreto, accediendo a través de una interfaz web desde cualquier lugar. SCAYLE dispone, por tanto, de una aplicación única que ofrece a usuarios remotos y que ha sido diseñada para aprovechar y rentabilizar el tiempo de computación residual.

Se pretende, por lo tanto, desarrollar una versión del programa de resolución ab Initio de estructuras macromoleculares Arcimboldo, creando un entorno web 2.0, plataforma sencilla de utilizar en remoto, para que los usuarios utilicen un modelo RaaS (Results as a Service), y así poder solucionar un problema concreto de un usuario concreto en un periodo breve de tiempo, sin necesidad de formación específica en supercomputación.

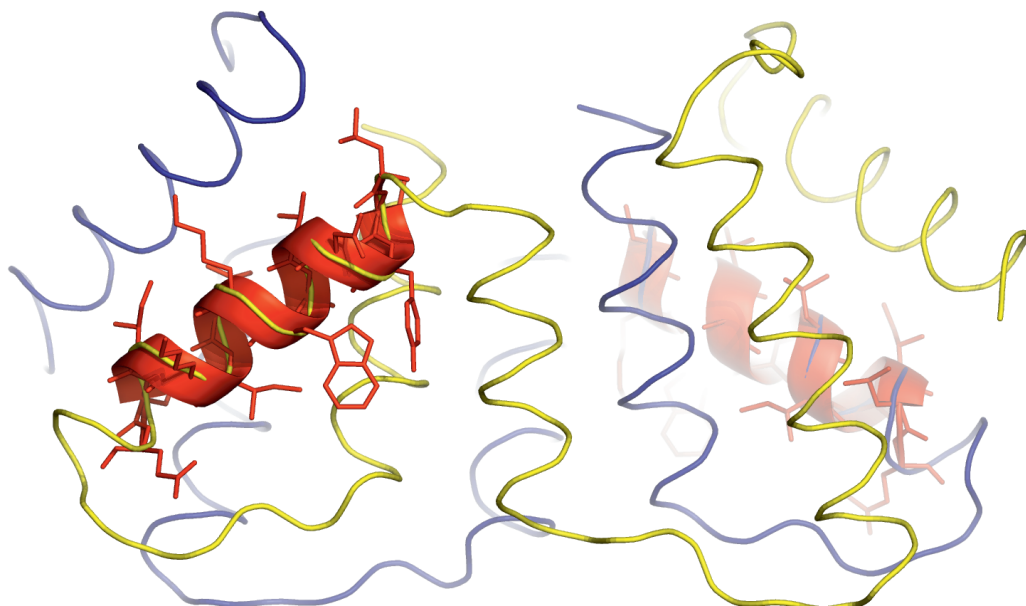
Desde su publicación y difusión en septiembre del año 2009, el programa Arcimboldo ha sido licenciado por más de 350 grupos de investigación en todo el mundo.

El proyecto se desarrolla como cooperación entre la empresa Catón sistemas alternativos, el grupo de la Dra. Usón en el IBMB-CSIC y SCAYLE con financiación del programa Industria de la Ciencia del CDTI.

En la mayoría de los casos, el sentido de la vista es el dominante para los humanos. Es, por tanto, fácil comprender que la visualización de las moléculas que juegan papeles fundamentales en los procesos biológicos que subyacen la medicina y la biotecnología juega un papel fundamental a la hora de estructurar todo nuestro conocimiento de tales sistemas. Entendemos "viendo", desde enfermedades hasta procesos industriales.

La cristalografía macromolecular proporciona una

visión de la estructura tridimensional de las macromoléculas biológicas con un altísimo grado de detalle y precisión a partir de datos de difracción medidos en sincrotrón (tales como el acelerador de partículas ALBA, en Barcelona). Sin embargo, esta técnica no proporciona una imagen directamente sino que es necesario establecer un modelo inicial, recurriendo a costosos experimentos que pueden durar meses, o bien por medio de computación exhaustiva.



Estructura de PRD2 y hélices superpuestas con cadenas laterales modeladas en las conformaciones standard, tal y como estaban colocadas en la solución correcta.

Desde septiembre de 2009, cuando Arcimboldo se puso a disposición de cualquier usuario académico, se han ido incorporando nuevas fuentes de información experimental o estereoquímica al método con el objetivo de volverlo más efectivo, lo cual ha redundado en un proyecto cada vez más ambicioso que requeriría una mayor fuente de cálculo.

La colaboración con SCAYLE ha sido clave a la hora de desarrollar nuevas técnicas y explotar nuevos métodos, de manera que en un futuro muy inmediato los usuarios de cristalografía no necesitarán disponer de los costosos recursos de hardware que requiere el método porque dispondrán de una pasarela web, sencilla e intuitiva, que les ayudará en el proceso de configuración de sus problemas e hipótesis y les permitirá ejecutar sus trabajos en remoto utilizando los recursos disponibles en Caléndula.

El éxito de Arcimboldo se basa en la correcta localización de fragmentos modelo tan pequeños como hélices de polialaninas de 10-14 residuos y se ha conseguido demostrar que, una vez colocados correctamente, se puede resolver la totalidad de la estructura, partiendo de un escaso 12% del total de la misma, colocado correctamente y aplicando los métodos de modificación de la densidad.

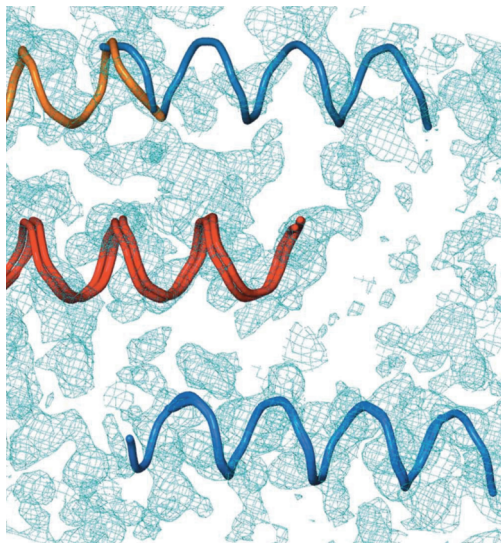
Las estructuras helicoidales suelen responder mejor al método. Usualmente las hélices contenidas en las proteínas suelen ser más largas que los 10-14 aminoácidos mencionados anteriormente, estos fragmentos modelo son rotados, trasladados y refinados utilizando el programa PHASER y el correspondiente **dataset** de reflexiones que nos proporcionan los usuarios. Colocar una hélice de 14 aminoácidos en una macromolécula que contenga una hélice de 20 residuos, puede conducir a la correcta ubicación del modelo 7 veces, con la única diferencia de que estarán desplazadas en el espacio sólo por un aminoácido.

Esto implica que el método demande una alta capacidad de cálculo, porque no sólo se encontrarán los fragmentos modelo correctamente colocados, si no todos aquellos que cumplan con ciertos requisitos de corte, de los cuales muchos no servirán como hipótesis inicial, pero que tienen que ser analizados por Arcimboldo, que decidirá "inteligentemente" con cuáles seguir adelante para construir el resto de la estructura. Todo esto implica que el método esté desarrollado en un entorno multisolución, donde se paralelizan las tareas en un grid de CONDOR, optimizando así los recursos disponibles y acortando muchísimo los tiempos de cálculo, lo cual encaja perfectamente con la

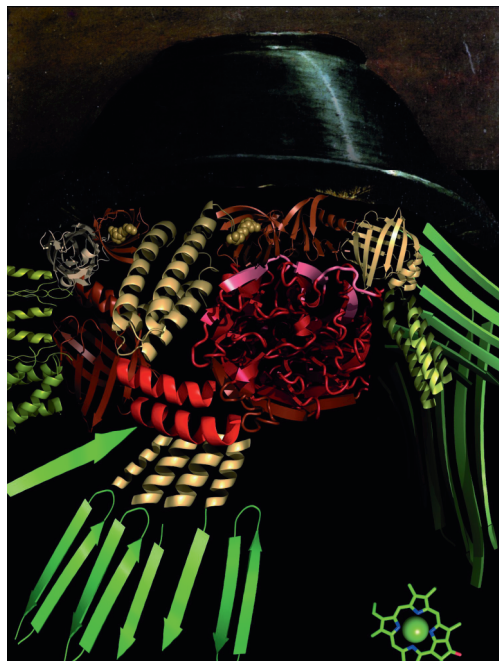
arquitectura disponible en SCAYLE, que es una herramienta informática estupenda para resolver este tipo de problemas, pero que constituye también una herramienta bastante compleja para el usuario de cristalografía promedio.

Al establecer este proyecto conjunto se aprovechan las grandes capacidades de cálculo y paralelización que ofrece Caléndula, que conjuntamente con una interfaz web que configure el problema y lo conecte con Arcimbolito constituirá una sólida herramienta de cálculo masivo en paralelo para los usuarios de biología molecular sin que éstos necesiten tener conocimientos de supercomputación.

Durante el período de colaboración con SCAYLE, y aprovechando la gran capacidad de cálculo de que ahora se dispone, se han añadido nuevas características para enriquecer el método, como por ejemplo, la búsqueda de fragmentos modelo un poco más sofisticados que las mencionadas hélices de polialaninas, que pueden obtenerse modelando las cadenas laterales o extrayéndolos directamente de estructuras homólogas. Para estos casos también se ha configurado Arcimbolito de manera que sea capaz de examinar un conjunto de posibles modelos de partida y seleccionar el más conveniente atendiendo a funciones de rotación y/o traslación sobre los mismos.



Tres subestructuras, de tres fragmentos cada una, que condujeron a la solución de la estructura PRD2 usando hélices modelo de 14 alaninas. Las hélices en rojo y naranja son comunes a las tres soluciones, la azul es ligeramente diferente y parcialmente superpuesta para dos de las soluciones y está fuera de la densidad para la tercera solución.



El sistema, bautizado con el nombre Arcimbolito en honor al pintor italiano Giuseppe Arcimbolito, ofrece imágenes reconocibles de proteínas a partir de unos fragmentos, de forma similar a los bodegones de vegetales del pintor en los que se pueden ver rostros humanos a partir de unos pocos elementos bien ubicados.

Participación de SCAYLE

SCAYLE dentro del marco del proyecto Arcimbolito tiene tres objetivos destacados:

- Integrar los sistemas con sus particularidades específicas y las de los subsistemas asociados como puedan ser los entornos paralelos, los sistemas de particiones, el subsistema de almacenamiento, el accounting, etc.
- Implementar los algoritmos de modelización dentro de los programas cristalográficos. Para ello tendrá que haber una generación y evaluación rápida de fragmentos modelo que faciliten una disminución del tiempo de búsqueda, manteniendo o aumentando el éxito global del procedimiento.
- Evaluar la viabilidad y eficiencia de los distintos algoritmos dentro de la capacidad del sistema de explotación multiusuario.

Además, la facilidad de uso y mantenimiento de una aplicación web soportada sobre una instalación externa como es SCAYLE hace que la distancia geográfica no suponga un obstáculo.

También se puede incorporar información experimental complementaria a la hipótesis estructural de múltiples maneras. Datos de difracción medidos en condiciones particulares, para provocar fenómenos físicos de resonancia pueden ser explotados al combinarlos con las hipótesis estructurales, así como cualquier condición de contorno que pueda ser expresada matemáticamente a los computadores. Estos trabajan de manera concertada, a través de un grid, eficiente y sofisticado.

En su otro extremo, la herramienta debe comunicarse con los usuarios, de la forma más cómoda, atractiva y sencilla posible. La interfaz web ayuda en la configuración de estrategias adecuadas dentro de los diferentes escenarios típicos. Será la puerta de entrada y la guía para resolver problemas de usuarios externos académicos e industriales. Se está asimismo poniendo a punto la explotación de esta herramienta compaginando usuarios simultáneos, lo que requiere un delicado equilibrio de prioridades.

Con Arcimboldo-Cloud se está desarrollando una solución a la falta de herramientas informáticas y a la complejidad del acceso por parte de los usuarios a los centros de supercomputación, ya que el desarrollo de software establecido para biología estructural se amplía a la explotación remota de cálculo masivo en paralelo. De esta forma la comunidad científica tendrá a su alcance una solución global de **software, middleware y hardware** a un problema concreto, accediendo a través de una interfaz web desde cualquier lugar. SCAYLE dispone, por tanto, de una aplicación única que ofrecer a usuarios remotos y que ha sido diseñada para aprovechar y rentabilizar el tiempo de computación residual. El año que viene comenzará a implicar usuarios seleccionados para poner a punto la apertura a la comunidad internacional.

Financiación del proyecto

Subprograma Industria de la Ciencia del CDTI.



Centro para el
Desarrollo
Tecnológico
Industrial

Código IDC-201001173



"Una manera de hacer Europa"

Curriculum vitae del investigador

Isabel Usón Finkenzeller es Profesora de Investigación ICREA adscrita al IBMB-CSIC (Instituto de Biología Molecular de Barcelona - Consejo Superior de Investigaciones Científicas).

Licenciada en Ciencias Químicas en 1987 por la Universidad de Zaragoza, y Dra. en Ciencias, Universidad de Zaragoza in 1992, Product Research Scientist en el European Research Center in Brussels de la multinacional Procter & Gamble durante 15 meses. Marie Curie HCM postdoc en la Universidad de Göttingen (Alemania) donde completó su Habilitación en la Facultad de Química (1996-2001) en el grupo del Prof. Sheldrick y continuó hasta 2003 como C2 (grupo emergente independiente).

Se incorporó en Septiembre 2003 al IBMB-CSIC como Profesora de Investigación ICREA.

Autora de 120 publicaciones científicas e investigadora principal de un grupo de 10 personas que ha desarrollado los programas cristalográficos Arcimboldo, Subix y Borges.