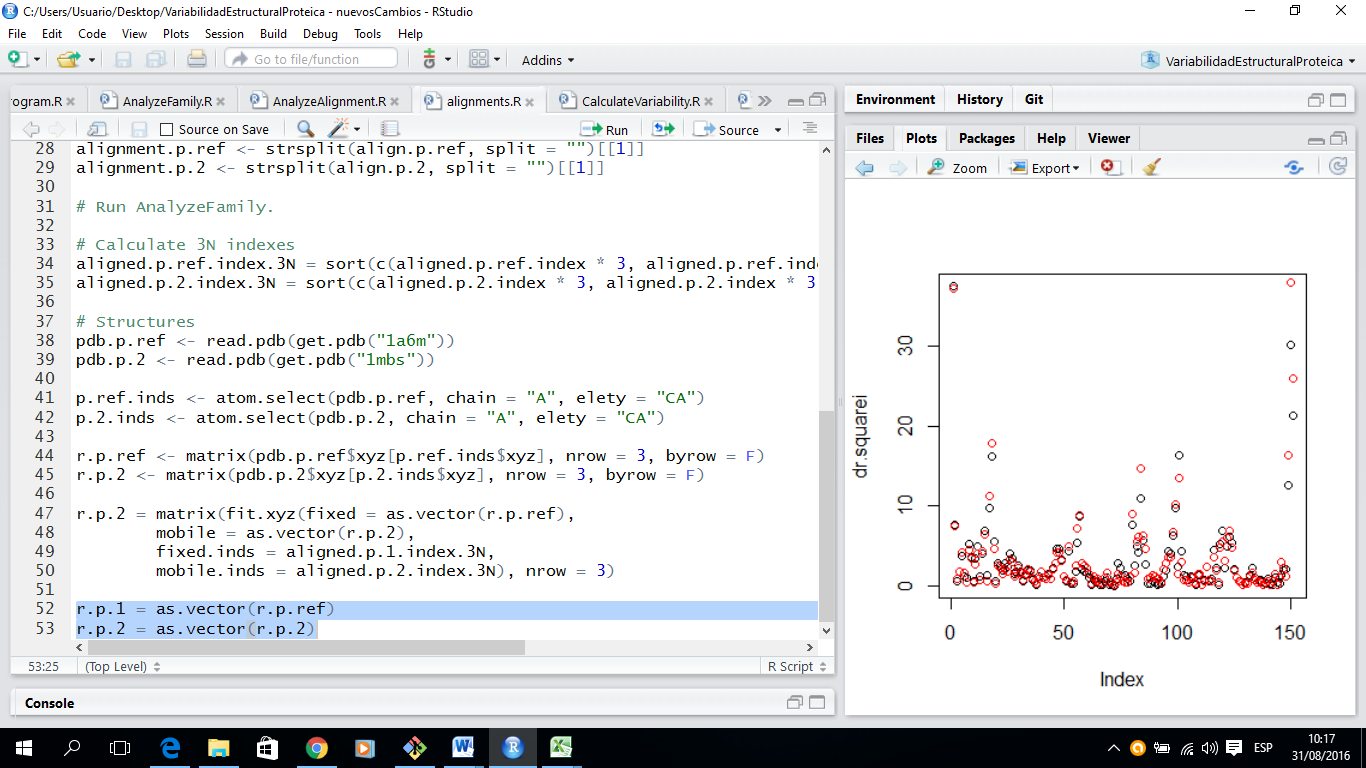
**COMPARACIÓN DE BLASTP PARA 1A6M CON EL CONJUNTO DE GLOBINAS DE HOMSTRAD**

* La mayor parte de las proteínas alineadas con alta identidad secuencial (aprox. 70 – 80 %) por BLASTP son mutantes ingenierizadas o proteínas iguales pero expuestas a Co o con otras variantes durante la cristalización. Mucha mioglobina de corazón de caballo con sus variantes. Si se desea hacer un dataset propio, hay que ver qué tan sencillo es eliminar secuencias indeseables con programas como CD-HIT.
* BLASTP encuentra a la mayor parte de las proteínas con alta %id del conjunto de Homstrad. Encuentra solo unas pocas con menor % id ya que no muestra resultados con %id <25 (aprox.).
* Usando psi-BLAST se encuentra a todas las proteínas del conjunto de Homstrad ya que este algoritmo permite encontrar proteínas más divergentes.

**ALINEAMIENTOS**

R permite alinear con MUSCLE con biocLite("muscle"). MUSCLE hace alineamiento secuencial, no estructural, pero para proteínas con alto % id (aprox. 80, lo que buscamos) obtengo el mismo alineamiento que da Homstrad. Más aún, después las roto usando la función de r fit.xyz() y obtengo resultados muy parecidos:



**Figura:**

Dr.square i para 1a6m vs 1mbs (%id = 83%).

Obtengo el mismo alineamiento para ambas.

Negro: alineadas con muscle y rotadas con fit.xyz

Rojo: alineadas por Homstrad y no re-rotadas.

También hay en R una función para alinear primero secuencialmente (usando MUSCLE) y después estructuralmente. La función es struct.aln(). Esta función no realiza alineamientos estructurales como los que buscamos, son más bien alineamientos secuenciales y luego se minimiza RMSD entre los índices obtenidos. Es decir que es como usar MUSCLE y fit.xyz.

**FAMILIAS DE PROTEÍNAS**

* **Azurin/Plastocianina. (OXIDORESUCTASA).**
* **Serin proteases. (HIDROLASA).**
* **Fosfolipasa (HIDROLASA).**
* Cistein proteasas. (HIDROLASA).
* Ferredoxinas. (OXIDOREDUCTASA).
* Glutatión S Trasferasa (TRANSFERASA).
* Lactato/malato deshidrogenasa (LIASA).
* Aspartic proteinasa. (HIDROLASA).
* Piridin Nucleótido disulfuro oxidoreductasa. (OXIDOREDUCTASA).
* Triosa fosfato isomerasa. (ISOMERASA).

**En función de su acción catalítica específica, las enzimas se clasifican en 6 grandes grupos o clases:**

* Clase 1: OXIDORREDUCTASAS. (Tengo 2)
* Clase 2: TRANSFERASAS. (Tengo 1)
* Clase 3: HIDROLASAS. (Tengo 4)
* Clase 4: LIASAS. (Tengo 1)
* Clase 5: ISOMERASAS. (Tengo 1)
* Clase 6: LIGASAS. (No tengo ninguna)

**COMPARACIONES ESTRUCTURALES**

Comparaciones en coordenadas cartesianas:

* Con rotación global.
* Con rotación local. Ventana en cada loop o región muy variable.
* Con rotación local. Entorno de empaquetamiento del sitio. R0 = 7.5/10.
* Similitud de entorno de cada sitio.

**COMPARACIONES DINÁMICAS**

* MSF
* nH
* nR

Programa de correlaciones múltiples

Archivos que agreguen columna con correlaciones a un archivo.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Family 1 | Family 2 | Family … |
| Exp – mut |  |  |  |
| Exp – strong |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Family 1 | Family 2 | Family … |
| CA |  |  |  |
| CM |  |  |  |

**ANÁLISIS CORE**

* Sitios sin gaps en el alineamiento estructural. No hay demasiados cambios.
* Sitios sin gaps y con vecinos sin gaps (6 o 3, me quedo sin sitios para analizar). Se mantienen las tendencias. Podría ser que algún sitio activo se conserve más… pero solo es marcado en globinas.

**CÓMO HACER MI PROPIO DATASET**

1. Obtener un representante de cada familia de proteínas deseada (por ejemplo, las 213 enzimas monoméricas de Yeh et al.).
2. Buscar homólogos cercanos con BLAST y/o lejanos con PSI-BLAST.
3. Remover secuencias redundantes usando CD-HIT (<http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit>).
4. Construir el alineamiento estructural múltiple con MUSCLE, CLUSTALW o similares para secuencias de alta similitud o con Dali, RAPIDO, MAMOOTH u otros para secuencias más divergentes.
5. Si se desea se puede construir un árbol filogenético y obtener las velocidades de evolución de los sitios con algunos de los algoritmos disponibles para generar árboles, como el “neighbour joining”, y mediante la inferencia Bayesiana y el modelo de evolución secuencial JTT.

A considerar: los árboles de especies pueden diferir a los de genes por transferencia horizontal de genes y/o duplicación.

**PROBLEMA CON SERIN PROTEASAS CM (solucionado)**

Descripción del error: aparecen NaNs en K usando c(r.CA, r.CM).

Búsqueda del error:

* Cálculo de la matriz K: chequeada por contraste con la función hessian() de Bio3d.
* Valores no numéricos: Valores NaN solo cuando agrego CM o cuando uso los CMs solos.
* Inspección CMs: índices repetidos en el pdb llevan a CMs iguales en 3 pares de sitios. Corregido en el programa de cálculo de CMs.

AZURIN/PLASTOCIANINAS

La plastocianina es una [cuproproteína](https://es.wikipedia.org/wiki/Cuproprote%C3%ADna" \o "Cuproproteína) involucrada en la [**cadena de transporte de electrones**](https://es.wikipedia.org/wiki/Cadena_de_transporte_de_electrones). Es una proteína **monomérica** con un [peso molecular](https://es.wikipedia.org/wiki/Peso_molecular) de alrededor de 10,5 [KDa](https://es.wikipedia.org/wiki/Dalton_(unidad)" \o "Dalton (unidad)) y 99 [aminoácidos](https://es.wikipedia.org/wiki/Amino%C3%A1cido) que se encuentra en la mayoría de las [plantas](https://es.wikipedia.org/wiki/Plantae), su nombre se debe a que se localiza en los cloroplastos y por su color azul en la forma oxidada. Las azurinas son proteínas relacionadas a las plastocianinas y se encuentran en bacterias.

SERIN PROTEASAS

Las serina proteasas son [hidrolasas](https://es.wikipedia.org/wiki/Hidrolasa) que degradan [enlaces peptídicos](https://es.wikipedia.org/wiki/Enlace_pept%C3%ADdico) de [péptidos](https://es.wikipedia.org/wiki/P%C3%A9ptido) y [proteínas](https://es.wikipedia.org/wiki/Prote%C3%ADna) y que poseen en su centro activo un [aminoácido](https://es.wikipedia.org/wiki/Amino%C3%A1cido) [serina](https://es.wikipedia.org/wiki/Serina" \o "Serina) esencial para la [catálisis enzimática](https://es.wikipedia.org/wiki/Cat%C3%A1lisis_enzim%C3%A1tica). Esta clase de enzimas (clasificadas como [EC](https://es.wikipedia.org/wiki/N%C3%BAmero_EC)[3.4.21](http://enzyme.expasy.org/EC/3.4.21)) incluye a la [tripsina](https://es.wikipedia.org/wiki/Tripsina), [quimotripsina](https://es.wikipedia.org/wiki/Quimotripsina" \o "Quimotripsina), [subtilisina](https://es.wikipedia.org/wiki/Subtilisina" \o "Subtilisina) y otras.

Las serina proteasas cortan la cadena polipeptídica en el lado [carboxilo](https://es.wikipedia.org/wiki/Carboxilo) de aminoácidos específicos, esto es, reconocen secuencias en su la [estructura primaria](https://es.wikipedia.org/wiki/Estructura_de_las_prote%C3%ADnas). Por ejemplo, la tripsina corta en el lado carboxilato de los residuos básicos como la [lisina](https://es.wikipedia.org/wiki/Lisina) o la [arginina](https://es.wikipedia.org/wiki/Arginina), mientras que la [quimiotripsina](https://es.wikipedia.org/wiki/Quimiotripsina" \o "Quimiotripsina) lo hace junto a residuos hidrófobos, como la [fenilalanina](https://es.wikipedia.org/wiki/Fenilalanina).

Existe una cierta homogeneidad estructural en cuanto a conformación estructural de las serina proteasas, lo que sugiere una relación evolutiva entre ellas. Por ejemplo, **las serina proteasas siempre poseen una**[**tríada catalítica**](https://es.wikipedia.org/wiki/Tr%C3%ADada_catal%C3%ADtica)**de [aspartato](https://es.wikipedia.org/wiki/Aspartato" \o "Aspartato),**[**histidina**](https://es.wikipedia.org/wiki/Histidina)**y serina ubicadas en la grieta catalíca del**[**sitio activo**](https://es.wikipedia.org/wiki/Sitio_activo). Además, estas **enzimas siempre poseen un «bolsillo» situado cerca de la serina del lugar activo**. En el caso de la [tripsina](https://es.wikipedia.org/wiki/Tripsina), este bolsillo le permite captar y mantener aminoácidos básicos (cargados positivamente), gracias a que la enzima posee en ese lugar el grupo carboxilo de la cadena lateral de un ácido aspártico; **es decir, la presencia de un residuo ácido en el bolsillo es el que confiere su especificidad a la tripsina, la afinidad hacia el corte en aminoácidos básicos de las proteínas diana, aminoácidos que interactúan mediante**[**interacción electrostática**](https://es.wikipedia.org/wiki/Enlace_i%C3%B3nico)**con el grupo ácido del bolsillo.**

FOSFOLIPASAS

Las fosfolipasas son una clase de [enzimas](https://es.wikipedia.org/wiki/Enzima) que [hidrolizan](https://es.wikipedia.org/wiki/Hidr%C3%B3lisis) los enlaces éster presentes en los [fosfolípidos](https://es.wikipedia.org/wiki/Fosfol%C3%ADpidos).