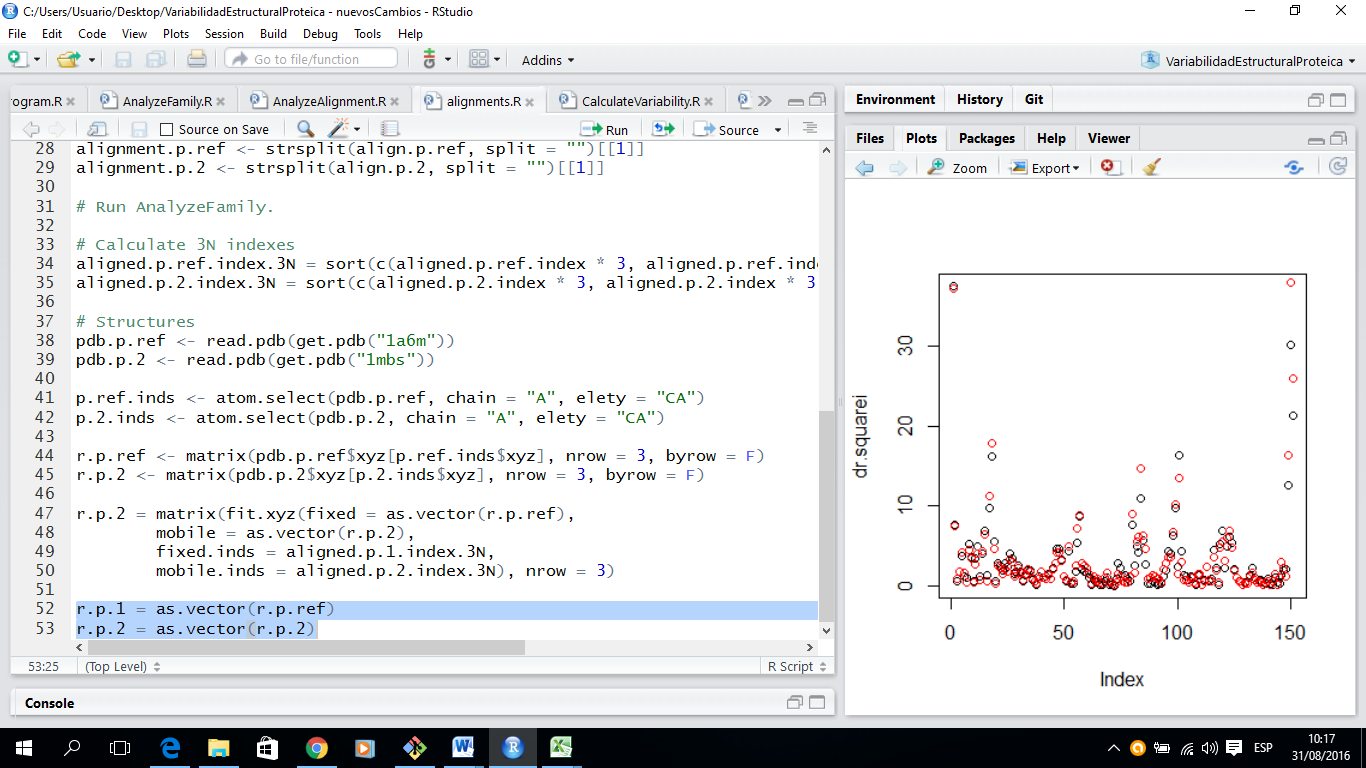
**COMPARACIÓN DE BLASTP PARA 1A6M CON EL CONJUNTO DE GLOBINAS DE HOMSTRAD**

* 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(). No me funciona en mi R. Sin embargo, no sé hasta qué punto hace alinemientos estructurales.

**FAMILIAS DE PROTEÍNAS**

* Plastocianina. (OXIDORESUCTASA). Falta p.ref, CM, correr y análisis
* **Cistein proteasas. (HIDROLASA).**
* **Serin proteases. (HIDROLASA).**
* **Fosfolipasa (HIDROLASA).**
* Ferredoxinas. (OXIDOREDUCTASA). FALTA ANÁLISIS
* Glutatión S Trasferasa (TRANSFERASA). FALTA ref CMS Y CORRER. FALTA ANÁLISIS
* Lactato/malato deshidrogenasa (LIASA). Corriendo. FALTA ANÁLISIS
* 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.