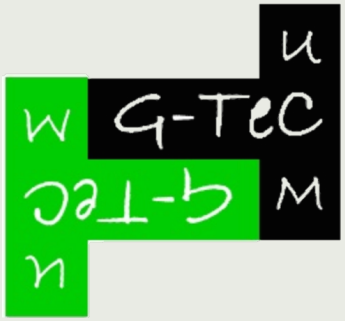
APLICACIÓN DE TÉCNICAS DE OPTIMIZACIÓN Y BIG DATA AL PROBLEMA DE BÚSQUEDA DE HOMOLOGÍAS EN BASES DE DATOS BIOLÓGICAS

Gabriel Antonio Valverde Castilla
Dra. Beatriz González-Pérez
Dra. Victoria López López

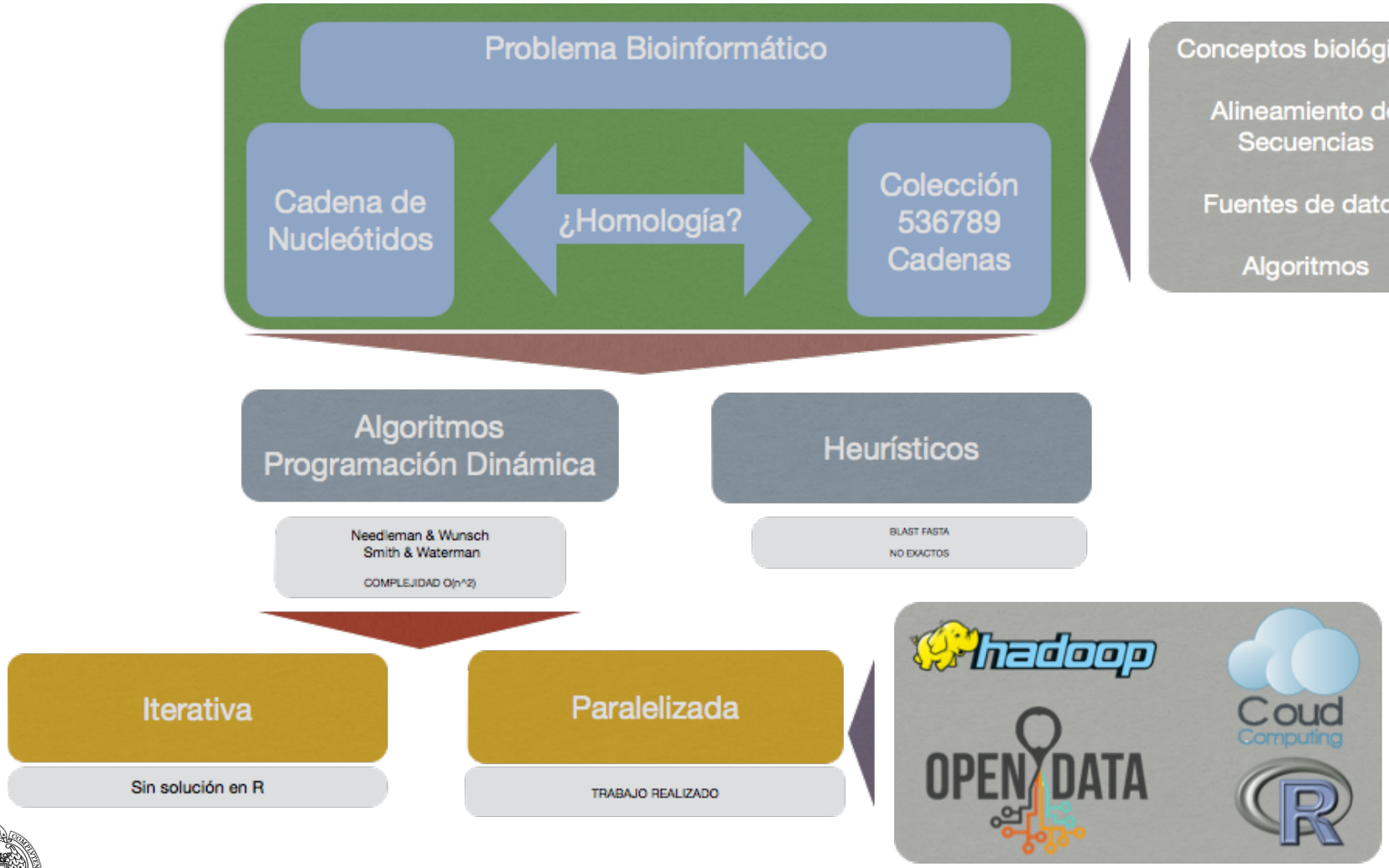


VII Jornadas de Usuarios de R
5 y 6 de Noviembre de 2015
Salamanca

CONTENIDO

- Motivación y objetivos
- Bases de datos
- Alineamiento de Secuencias
- Algoritmos
- Implementación
- Resultados

Motivación


Problema Bioinformático
Cadena de Nucleótidos
¿Homología?
Colección 536789 Cadenas
Conceptos biológicos
Alineamiento de Secuencias
Fuentes de datos
Algoritmos
Algoritmos Programación Dinámica
Needleman & Wunsch
Smith & Waterman
COMPLEJIDAD O(n^2)
Heurísticos
BLAST FASTA
NO EXACTOS
Iterativa
Sin solución en R
Paralelizada
TRABAJO REALIZADO
hadoop
Coud Computing
OPEN DATA
R

Paquetes de R

Paquete	Descripción	Funciones de interés
seqinr	Análisis exploratorio y visualización de datos de secuencias biológicas.	Read.fasta, write.fasta, s2c, c2s, dotplot, query…
Biostrings	Algoritmos de alineamiento y utilidades para la manipulación eficiente de largas secuencias biológicas.	readFASTA, writeFASTA, pairwiseAlignment, getTrans…
ff	Estructuras de almacenamiento masivo de datos.	Ff,ffsave,ffload,open.ff
Data.table	Versión mejorada de los data.frame de R. Con algoritmos de consultas más veloces basados en B-tree.	All.equal, :=, between, in, chmatch, duplicated, fread, melt.data.table, rbindlist …
PLYR	Como data.table nos permite hacer consultas más veloces sobre data.frame. Basándose en Split + Combine.	Ddply,llply.
doParallel	Permite la paralización multicores de máquina de algoritmos.	Mclapply, registerDoParalll, stopImplicitCluster.
RHADOOP	Permite la interacción ente R y Hadoop, basado en el modelo de programación MapReduce	Rmr, RHDFS,RBASE,PLYRR Mapreduces,hdfs.init,hdfs.put,hdfs.get,…

Bases De Datos

- NoSql
  - Nuevo sistema de bases de datos más genérico que los convencionales buscando rendimiento, velocidad y generalidad.
- Genómica
  - Pioneros en las bases de datos abiertas concienciados con su valor. NCBI EBI. Genbank 28.000 millones de pares de bases (nucleótidos), correspondientes a más de 22 millones de secuencias
- Fasta
  - Formato genérico para tratar con datos de secuencias de nucleótidos y aminoácidos que hemos utilizado.

FASTA

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

```
> choosebank("genbank")
> ccnd3hs <- query("ccnd3hs","sp=homo sapiens AND k=ccnd3@")
> sapply(ccnd3hs$req,getName)
[1] "AF517525.CCND3"   "BC011616.CCND3"   "CR542246"         "HUMCCND3A.CCND3"
[5] "HUMCCND3PS.PE1"   "HUMCCNDB04.CCND3" "HUMCYCD3A.CCND3"
> sapply(ccnd3hs$req,getLength)
[1] 879 879 879 879 537 559 879
> getSequence(ccnd3hs$req[[2]])[1:30]
 [1] "a" "t" "g" "g" "a" "g" "c" "t" "g" "c" "t" "g" "t" "g" "t" "t" "g" "c" "g" "a" "a"
[22] "g" "g" "c" "a" "c" "c" "c" "g" "g"
> getTrans(ccnd3hs$req[[2]])[1:30]
 [1] "M" "E" "L" "L" "C" "C" "E" "G" "T" "R" "H" "A" "P" "R" "A" "G" "P" "D" "P" "R" "L"
[22] "L" "G" "D" "Q" "R" "V" "L" "Q" "S"
```

- Biostrings
- Seqirn

Alineamiento de Secuencias

SECUENCIAS DE AMINOACIDOS

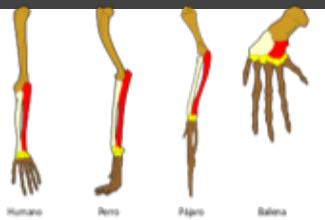
Cadena lineal, finita y ordenada de símbolos de un alfabeto que en este caso determinan la funcionalidad de una proteína.

Maxima Coincidencia de Carácteres

mach
imach
gap

Algoritmos de Programación Dinámica

Similaridad



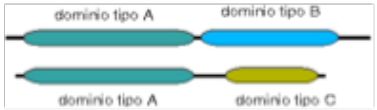
Homología

Concepto biológico
Ancestros comunes

Alineamiento Global

Algoritmo de Needleman-Wunsch

Secuencias parecidas, de longitud semejante
Intenta alinear todos los elementos



Alineamiento Global

Alineamiento Local

Alineamiento Local

Algoritmo *Smith-Waterman*

Divergencia estructural

Características puntuales similares

2 SECUENCIAS NO ALINEADAS

L	G	P	S	S	K	Q	T	G	K	G	S	S	R	I	W	D	N	
L	N	I	T	K	S	A	G	K	G	A	I	M	R	L	G	D	A	

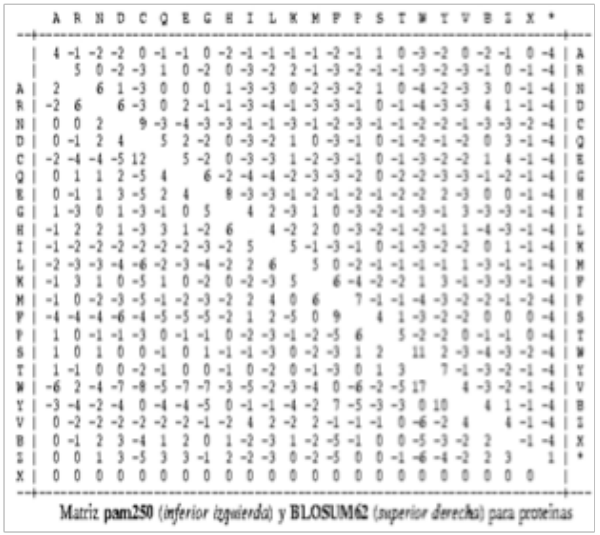
ALINEAMIENTO GLOBAL

L	G	P	S	S	K	Q	T	G	K	G	S	-	S	R	I	W	D	N
					•			•	•	•				•			•	
L	N	-	I	T	K	S	A	G	K	G	A	I	M	R	L	G	D	A

ALINEAMIENTO LOCAL

-	-	-	-	-	-	-	T	G	K	G	-	-	-	-	-	-	-	-
									•	•								
-	-	-	-	-	-	-	A	G	K	G	-	-	-	-	-	-	-	-

Matriz de Puntuación



Matriz **pam250** (*inferior izquierda*) y **BLOSUM62** (*superior derecha*) para proteínas

Algoritmo: Needleman - Wunsch

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-41	-49	-57	-55	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-26	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-5	3	-5
A	-48	-80	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

$t_{ij} = \max(t_{i-1,j-1} + s(x_i, y_j), t_{i,j-1} + d, t_{i,j-1} + d)$

$s(x_i, y_j)$ = valor matriz de sustitución Blosum62 en nuestro caso

$t_{23} = \max(t_{12} + s(x_P, y_E), t_{2,2} + d, t_{2,3} + d)$
$t_{23} = \max(-8 + 1, -2 - 8, -16 - 8) = -9$

H E A G AW GHE_E
_ _ A _ AW _HEAE

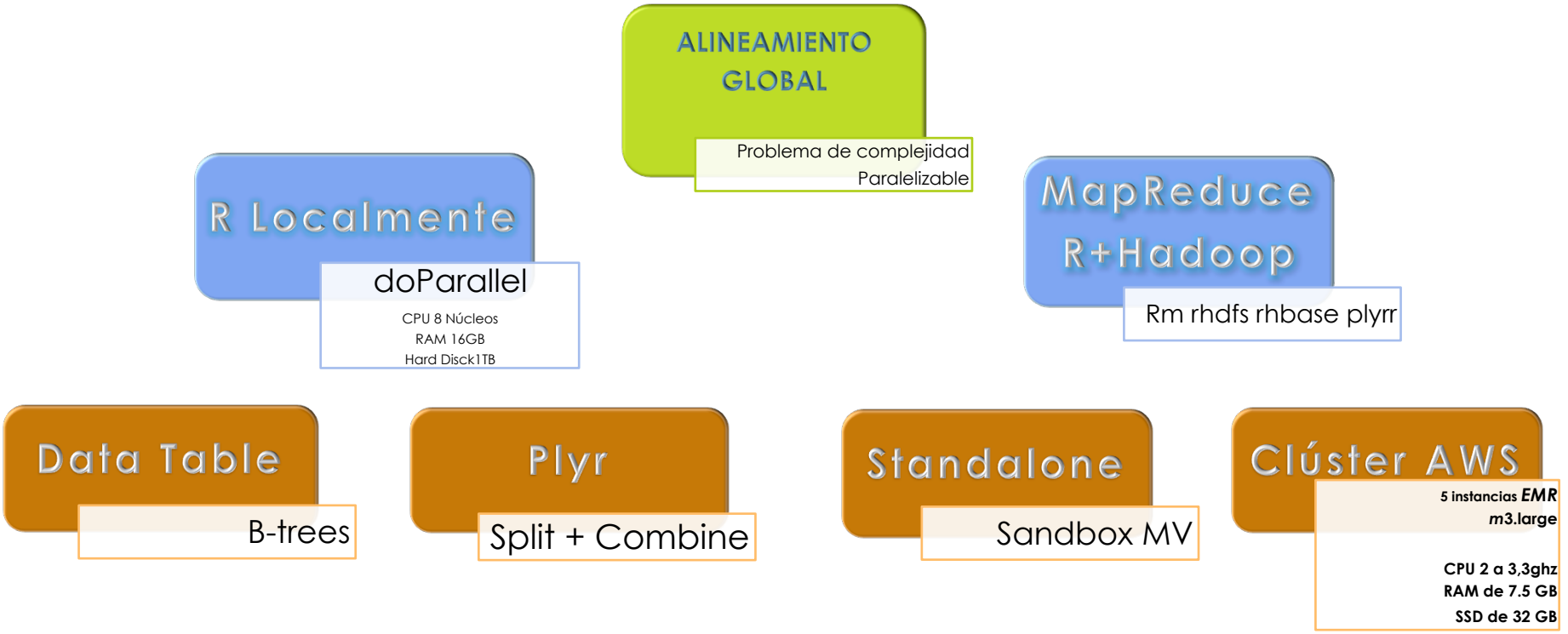
H E A G A W G H E _ E
_ P A _ _ W _ H E A E

Al realizar alineamiento de secuencias hemos de tener en cuenta:

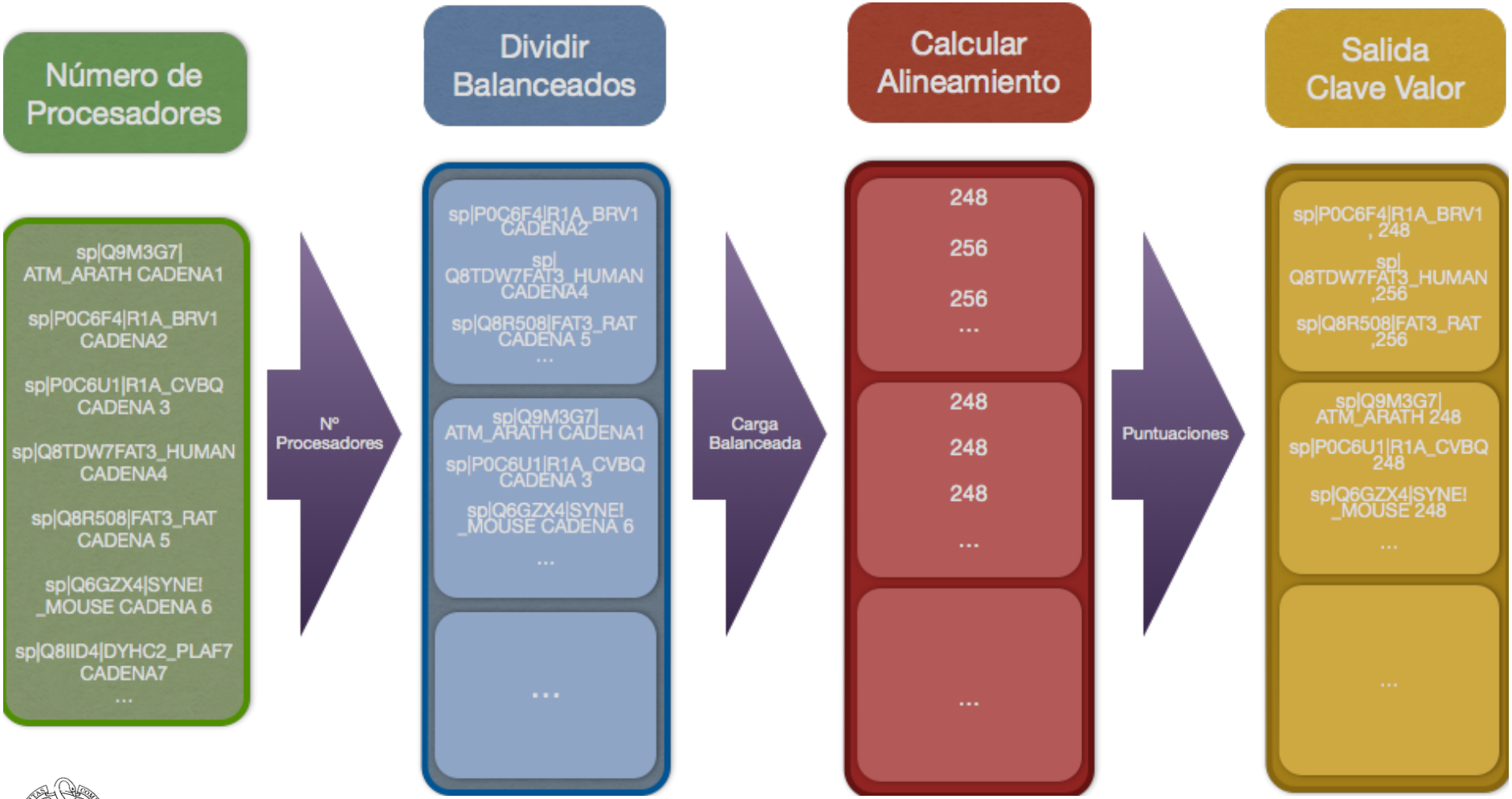
- Tipo de alineamiento
- ADN (más sensible) o traducción
- Elegir un sistema de puntuaciones adecuado. PAM BLOSUM.
- Métodos estadísticos para explicarlo.

http://ubio.bioinfo.cnio.es/people/fabascal/Teaching/Alineamiento_secuencias/teoria.html

Implementación



Método I: R Localmente

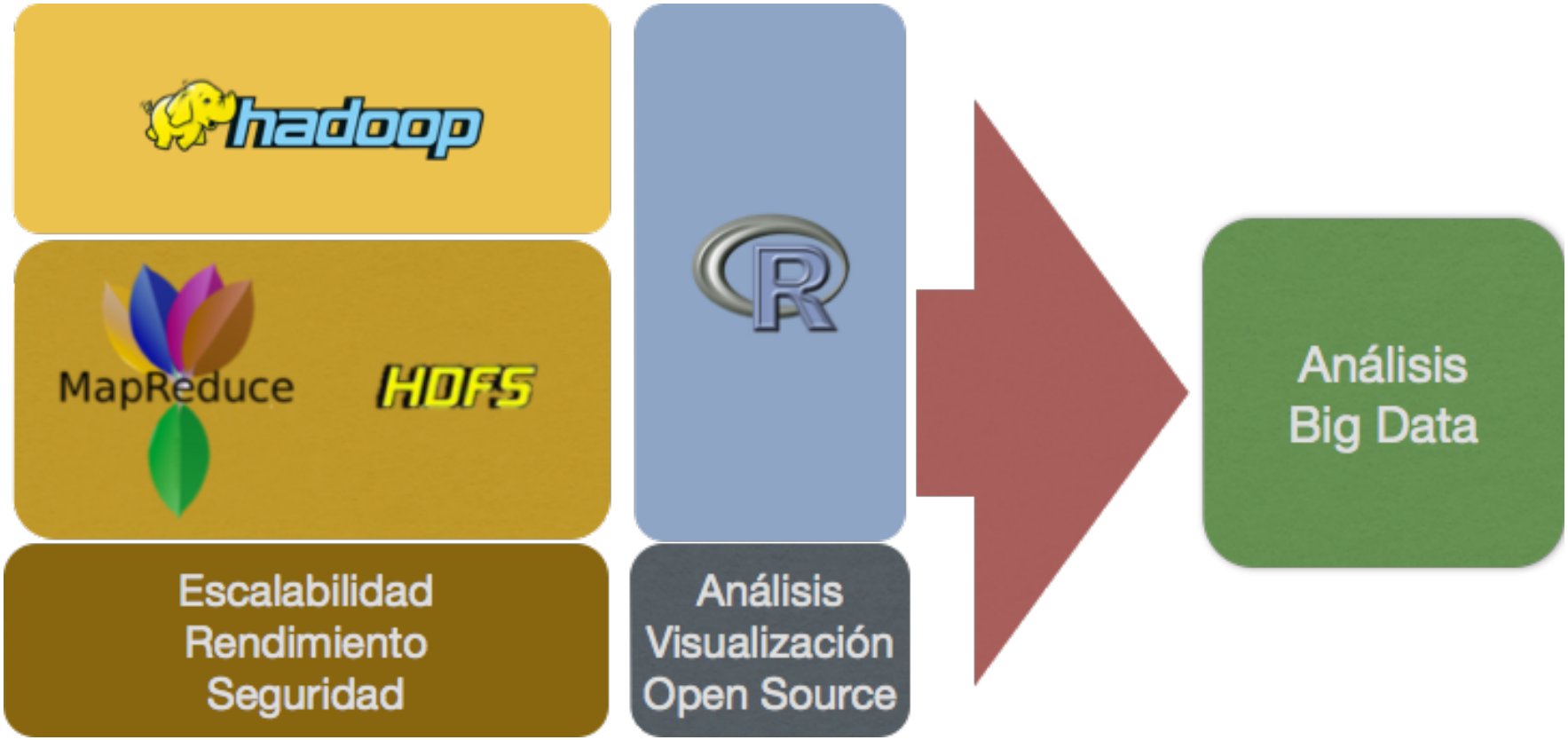


Método I: R Localmente

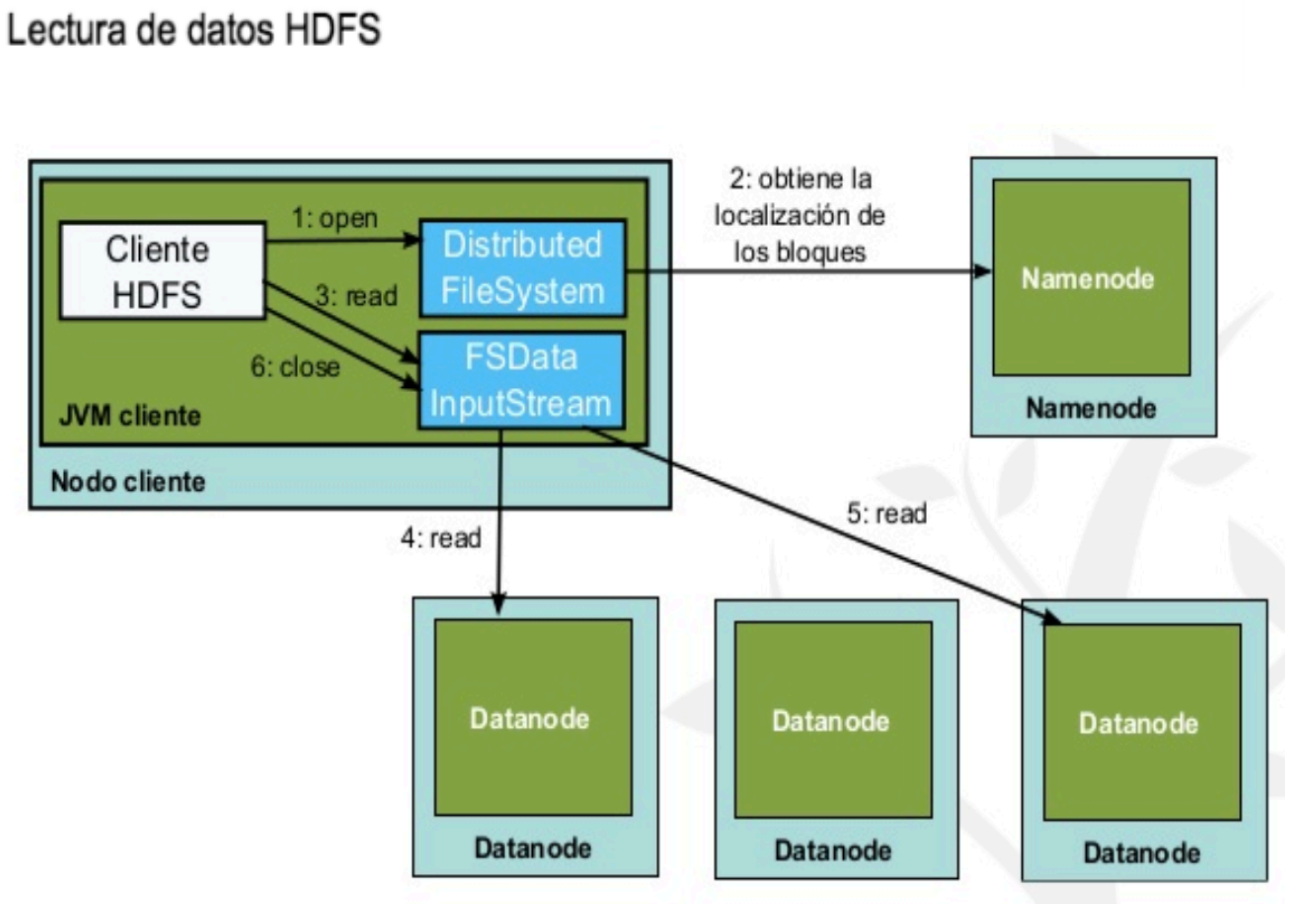
- Librerías de R utilizadas:

PAQUETE DE R	DEFINICION	FUNCIONES
Data.Table [14R]	Paquete que mediante el uso de árboles binarios y otras forma de optimización en el uso de índices, ha mejorado el rendimiento de los objetos data.frame. Nos permite hacer operaciones típicas de las bases de datos, como son select, where, by, y crear nuevas columnas de forma ágil.	As.data.table: Convierte en data.table otros objetos de r. .SD: permite aplicar una función a varias columnas de un data.table. setkey: Permite definir una columna como clave para realizar operaciones por grupo.
Plyr	Facilita el cálculo vectorial y también consta de distintas funciones típicas de la gestión de bases de datos. Existe una versión para trabajar con Hadoop.	Ddply, permite crear nuevas variables a partir de las variables que forman el conjunto de datos.
parallel	Nos permite realizar paralelización local en el número de nodos de forma transparente en R.	Foreach, aplica un for sobre el conjunto definido, fragmentando el conjunto de datos en memoria para establecer acceso distribuido en función de los nodos.

MétodoII : MapReduce R + Hadoop



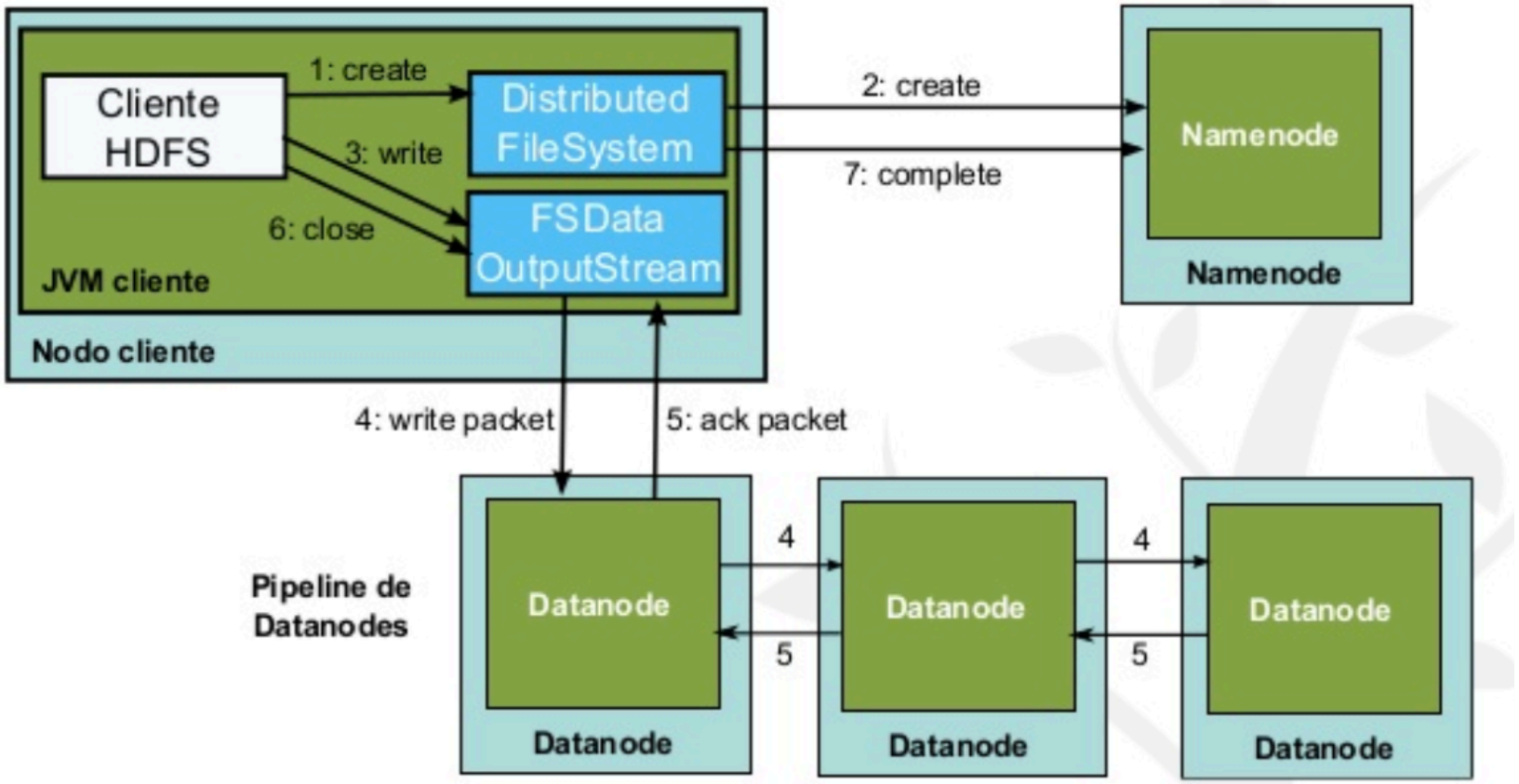
HDFS



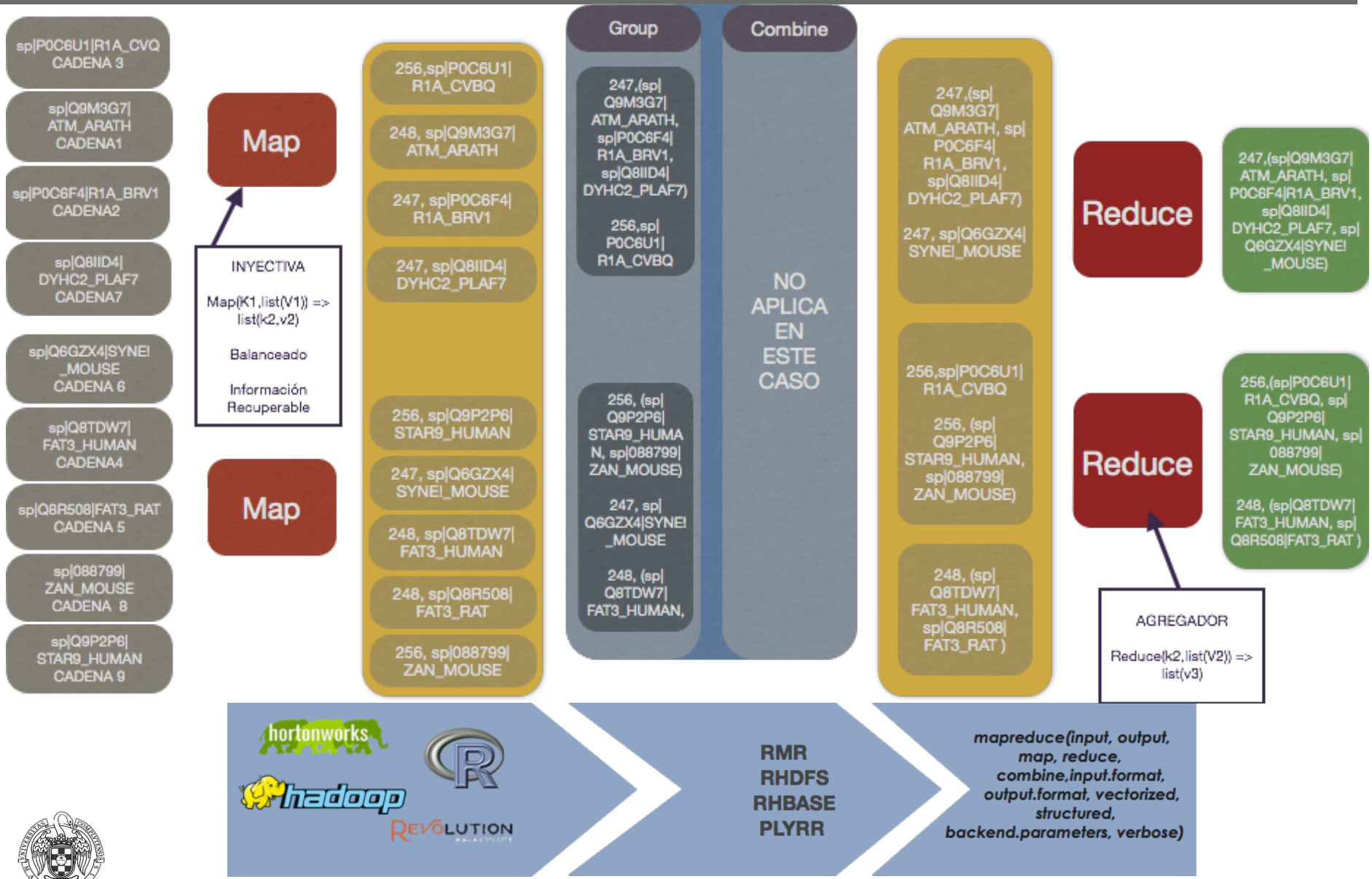
- Sistema de distribución de archivos de Hadoop. Ventajas:
  - Crecimiento horizontal vs problemas de espacio.
  - Replicas vs Fallos del sistema.
  - Distribución vs Tiempo de acceso
  - Paralelización vs Tiempo de ejecución
- Inconvenientes:
  - Elevada latencia con pequeños datos
  - Cambios se introducen siempre al final
  - No se permite escritura simultanea sobre ficheros

HDFS

Escritura de datos HDFS

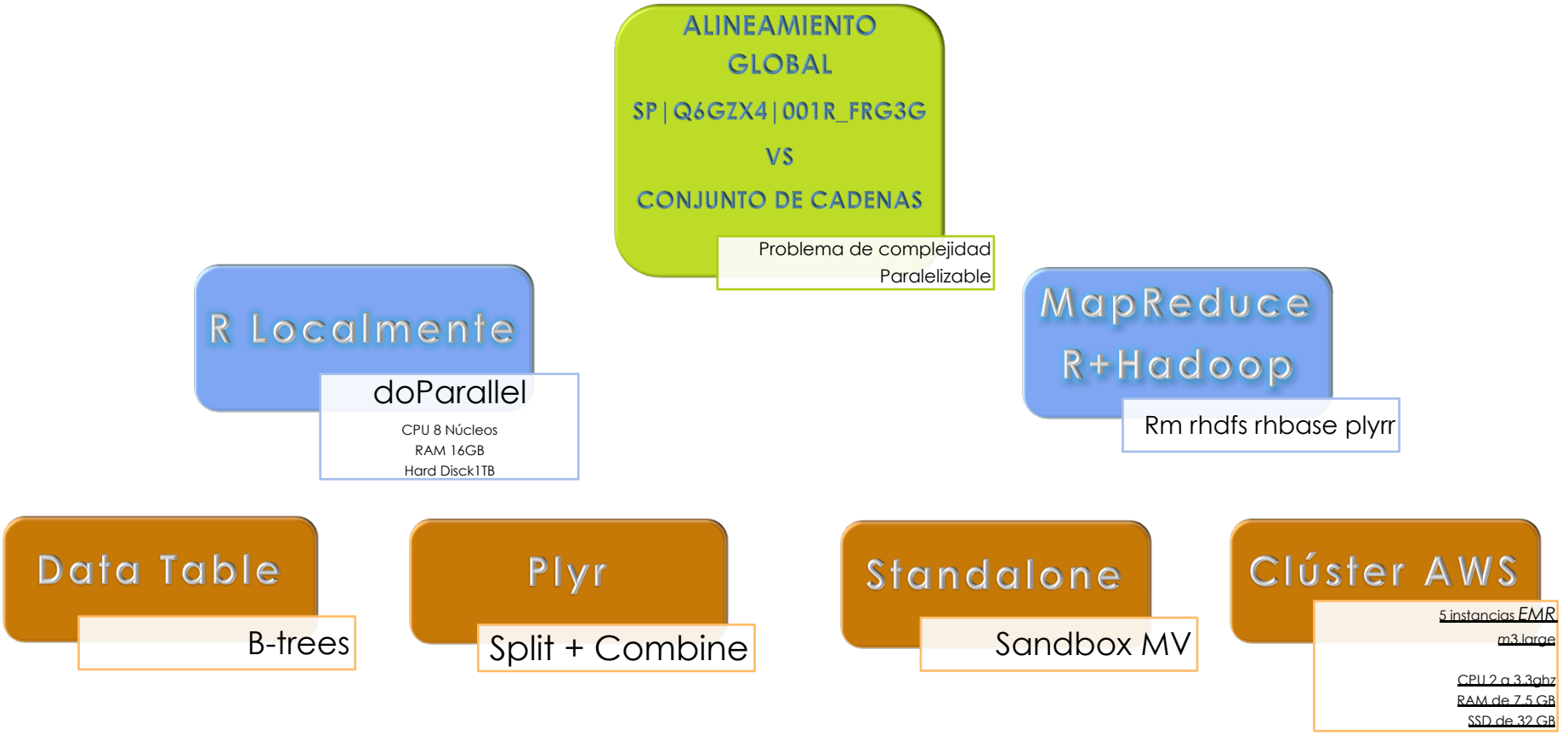


MapReduce


sp|P0C6U1|R1A_CVQ CADENA 3
sp|Q9M3G7| ATM_ARATH CADENA1
sp|P0C6F4|R1A_BRV1 CADENA2
sp|Q8IID4| DYHC2_PLAF7 CADENA7
sp|Q6GZX4|SYNE! _MOUSE CADENA 6
sp|Q8TDW7| FAT3_HUMAN CADENA4
sp|Q8R508|FAT3_RAT CADENA 5
sp|088799| ZAN_MOUSE CADENA 8
sp|Q9P2P6| STAR9_HUMAN CADENA 9
Map
Map
INYECTIVA
Map(K1,list(V1)) => list(k2,v2)
Balanceado
Información Recuperable
256,sp|P0C6U1| R1A_CVBQ
248, sp|Q9M3G7| ATM_ARATH
247, sp|P0C6F4| R1A_BRV1
247, sp|Q8IID4| DYHC2_PLAF7
256, sp|Q9P2P6| STAR9_HUMAN
247, sp|Q6GZX4| SYNE!_MOUSE
248, sp|Q8TDW7| FAT3_HUMAN
248, sp|Q8R508| FAT3_RAT
256, sp|088799| ZAN_MOUSE
Group
Combine
247,(sp| Q9M3G7| ATM_ARATH, sp|P0C6F4| R1A_BRV1, sp|Q8IID4| DYHC2_PLAF7)
256,sp| P0C6U1| R1A_CVBQ
256, (sp| Q9P2P6| STAR9_HUMAN, sp|088799| ZAN_MOUSE)
247, sp| Q6GZX4|SYNE! _MOUSE
248, (sp| Q8TDW7| FAT3_HUMAN,
NO APLICA EN ESTE CASO
247,(sp| Q9M3G7| ATM_ARATH, sp| P0C6F4| R1A_BRV1, sp|Q8IID4| DYHC2_PLAF7)
247, sp|Q6GZX4| SYNE!_MOUSE
256,sp|P0C6U1| R1A_CVBQ
256, (sp| Q9P2P6| STAR9_HUMAN, sp|088799| ZAN_MOUSE)
248, (sp| Q8TDW7| FAT3_HUMAN, sp|Q8R508| FAT3_RAT)
Reduce
Reduce
247,(sp|Q9M3G7| ATM_ARATH, sp| P0C6F4|R1A_BRV1, sp|Q8IID4| DYHC2_PLAF7, sp| Q6GZX4|SYNE! _MOUSE)
256,(sp|P0C6U1| R1A_CVBQ, sp| Q9P2P6| STAR9_HUMAN, sp| 088799| ZAN_MOUSE)
248, (sp|Q8TDW7| FAT3_HUMAN, sp| Q8R508|FAT3_RAT)
AGREGADOR
Reduce(k2,list(V2)) => list(v3)
hortonworks
hadoop
REVOLUTION
RMR
RHDFS
RHBASE
PLYRR
mapreduce(input, output, map, reduce, combine,input.format, output.format, vectorized, structured, backend.parameters, verbose)

Nombre	Función
INPUT	Directorio de los datos de entrada que pueden provenir tanto de HDFS, incluidos por medio de la función to.dfs, como de la salida de otro mapreduce.
OUTPUT	Directorio de HDFS donde queremos que se almacenen, si es vacío puede ser enviado a un nuevo map, reduce, or from.dfs
MAP	Una función de R que recibe dos valores como entrada, clave y valor. Devuelve con la misma estructura. keyval(k, v, vectorized = FALSE) Es su principal aliada, es la función que ayuda a definir estás parejas.
Reduce	Función que se aplica a la salida del map y devuelve agrupado los resultados por claves.
Make.input/output.format	Permite crear los formatos de los datos de entrada,
Imput/output.format	Permite definidor los formatos de forma que puedan ser transformados al único que soporta Hadoop Streaming.
Vectorized	Permite procesar múltiples registros al mismo tiempo, en lugar de línea a línea.
backend	Parámetros adicionales de Hadoop, como son las limitaciones de memoria, el número de tareas por nodo,… Nos permite seleccionar trabajar en local para hacer debug de código.
To.map /to.reduce	Convierte cualquier función en una función apta para aplicar en este proceso. Incluyendo una función referente a value y otra a key.

Escenario



http://www.uniprot.org/uniprot/Q6GZX4?version=*

Salida Local

```
> head(Puntuaciones,9)
                    Names puntuaciones
1:  sp|P83570|GWA_SEPOF             2
2:  sp|P84761|ACI_TRIGI             3
3: sp|P83568|ILME_SEPOF             3
4: sp|P24272|LUXE_VIBFI             3
5:  sp|P62970|TRH_BOMOR             3
6:  sp|P62971|TRH_NOTVI             3
7:    sp|P62968|TRH_PIG             3
8:  sp|P62969|TRH_SHEEP             3
9: sp|P35904|ACH1_ACHFU             4
```

```
> tail(Puntuaciones,9)
                    Names puntuaciones
1: sp|Q8NF91|SYNE1_HUMAN           256
2: sp|Q6ZWR6|SYNE1_MOUSE           256
3: sp|Q8WXH0|SYNE2_HUMAN           256
4: sp|Q9I7U4|TITIN_DROME           256
5: sp|Q8WZ42|TITIN_HUMAN           256
6: sp|A2ASS6|TITIN_MOUSE           256
7:  sp|O30409|TYCC_BREPA           256
8:  sp|Q6GZX4|001R_FRG3G           256
9: sp|Q23551|UNC22_CAEEL           256
```

Salida RHadoop

	V1	Cadena
1:	247	sp|P25464|ACVS_ACRCH,sp|Q12955|ANK3_HUMAN,sp|Q9M3G7|ATM_ARATH,sp|Q8IID4|DYHC2_PLAF7,sp|P58107|EPIPL_HUMAN,sp|P97412|LYST_MOUSE,
2:	248	sp|Q20911|CUBN_CAEEL,sp|Q9SMH5|DYHC2_CHLRE,sp|O88277|FAT2_RAT,sp|Q8TDW7|FAT3_HUMAN,sp|Q8R508|FAT3_RAT,sp|Q6V0I7|FAT4_HUMAN,
3:	249	sp|Q9MBF8|DYH1B_CHLRE,sp|Q8BW94|DYH3_MOUSE,sp|Q9C0G6|DYH6_HUMAN,sp|Q19542|DYHC2_CAEEL,sp|Q27171|DYHC_PARTE,sp|P36022|DYHC_YEAST,
4:	250	sp|P04114|APOB_HUMAN,sp|Q96M86|DNHD1_HUMAN,sp|A7X2C3|EBHA_STAA1,sp|Q931R6|EBHA_STAAM,sp|Q99U54|EBHA_STAAN,sp|Q14517|FAT1_HUMAN,
5:	251	sp|Q19319|CADH4_CAEEL,sp|P0C6F1|DYH2_MOUSE,sp|Q63170|DYH7_RAT,sp|Q8IBG1|DYHC1_PLAF7,sp|Q9C1M7|DYHC_ASHGO,sp|P34036|DYHC_DICDI,
6:	252	sp|Q5SSE9|ABCAD_MOUSE,sp|Q7TMA5|APOB_RAT,sp|O88738|BIRC6_MOUSE,sp|Q69Z23|DYH17_MOUSE,sp|Q9P2D7|DYH1_HUMAN,sp|Q63164|DYH1_RAT,
7:	253	sp|Q9SRU2|BIG_ARATH,sp|Q9NR09|BIRC6_HUMAN,sp|Q8IVF4|DYH10_HUMAN,sp|Q96DT5|DYH11_HUMAN,sp|Q9UFH2|DYH17_HUMAN,sp|Q9SMH3|DYH1A_CHLRE,
8:	254	sp|Q86UQ4|ABCAD_HUMAN,sp|Q9N4M4|ANC1_CAEEL,sp|B9G2A8|BIG_ORYSJ,sp|Q8TE73|DYH5_HUMAN,sp|Q8VHE6|DYH5_MOUSE,sp|Q96JB1|DYH8_HUMAN,
9:	255	sp|Q91XQ0|DYH8_MOUSE,sp|A2ARZ3|FSIP2_MOUSE,sp|Q8VHN7|GPR98_MOUSE,sp|O95714|HERC2_HUMAN,sp|Q01886|HTS1_COCCA,sp|Q80W93|HYDIN_MOUSE,
10:	256	sp|Q6GZX4|001R_FRG3G,sp|C6KTB7|ALTH1_PLAF7,sp|O68006|BACA_BACLI,sp|O68008|BACC_BACLI,sp|Q6GZX4|001R_FRG3G,sp|Q54CU4|COLA_DICDI,

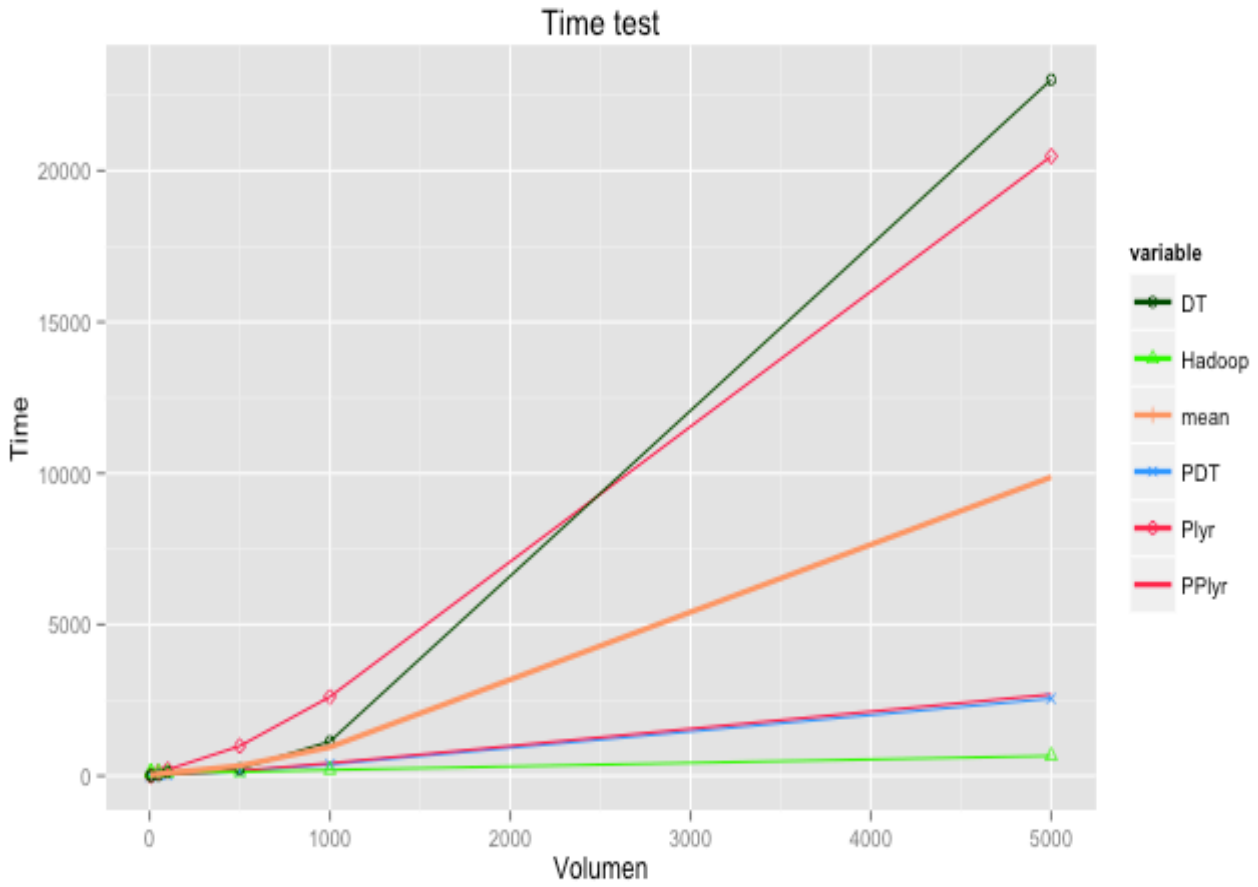
Salida RHadoop

```
[[1]]
 [1] "sp|P25464|ACVS_ACRCH"  "sp|Q12955|ANK3_HUMAN"  "sp|Q9M3G7|ATM_ARATH"   "sp|Q8IID4|DYHC2_PLAF7" "sp|P58107|EPIPL_HUMAN" "sp|P97412|LYST_MOUSE"
 [7] "sp|Q4WLW5|NRP12_ASPFU" "sp|Q4WAZ9|NRP14_ASPFU" "sp|P0C6F4|R1A_BRV1"    "sp|Q9YN02|RPOA_PRRS1"  "sp|Q8B912|RPOA_PRRSB"  "sp|Q8I8U7|TRA1_DROME"
[13] "sp|P38811|TRA1_YEAST"  "sp|Q80TY5|VP13B_MOUSE"

[[2]]
 [1] "sp|Q20911|CUBN_CAEEL"  "sp|Q9SMH5|DYHC2_CHLRE" "sp|O88277|FAT2_RAT"    "sp|Q8TDW7|FAT3_HUMAN"  "sp|Q8R508|FAT3_RAT"    "sp|Q6V0I7|FAT4_HUMAN"
 [7] "sp|Q8NEZ4|MLL3_HUMAN"  "sp|Q83034|POLG_RTSVA"  "sp|P78527|PRKDC_HUMAN" "sp|P0C6F6|R1A_BC512"   "sp|P0C6V6|R1A_PEDV7"   "sp|Q008X5|R1A_WBV24"
[13] "sp|Q04561|RPOA_PRRSL"  "sp|O43103|SID2_USTMA"  "sp|Q9Y4A5|TRRAP_HUMAN" "sp|Q4U4S6|XIRP2_MOUSE"

[[3]]
 [1] "sp|Q9MBF8|DYH1B_CHLRE" "sp|Q8BW94|DYH3_MOUSE"  "sp|Q9C0G6|DYH6_HUMAN"  "sp|Q19542|DYHC2_CAEEL" "sp|Q27171|DYHC_PARTE"  "sp|P36022|DYHC_YEAST"
 [7] "sp|A6QGY5|EBHB_STAAE"  "sp|Q9NYQ8|FAT2_HUMAN"  "sp|Q5F226|FAT2_MOUSE"  "sp|Q8BNA6|FAT3_MOUSE"  "sp|Q8NDA2|HMCN2_HUMAN" "sp|Q66431|L_DUGBA"
[13] "sp|Q8BRH4|MLL3_MOUSE"  "sp|Q7TPH6|MYCB2_MOUSE" "sp|Q86WI1|PKHL1_HUMAN" "sp|Q96662|POLG_BVDVC"  "sp|P19711|POLG_BVDVN"  "sp|Q9DEI1|PRKDC_XENLA"
[19] "sp|P0C6F7|R1A_BC133"   "sp|P0C6F5|R1A_BC279"   "sp|P0C6F8|R1A_BCHK3"   "sp|P0C6T7|R1A_BCRP3"   "sp|P0C6V1|R1A_CVMJH"   "sp|P0C6V2|R1A_CVPPU"
[25] "sp|P0C6V5|R1A_IBVM"    "sp|Q9WJB2|RPOA_PRRSR"  "sp|A0MD28|RPOA_PRRSS"

[[4]]
 [1] "sp|P04114|APOB_HUMAN"  "sp|Q96M86|DNHD1_HUMAN" "sp|A7X2C3|EBHA_STAA1"  "sp|Q931R6|EBHA_STAAM"  "sp|Q99U54|EBHA_STAAN"  "sp|Q14517|FAT1_HUMAN"
 [7] "sp|Q2PZL6|FAT4_MOUSE"  "sp|C6KTD2|HKNMT_PLAF7" "sp|Q9R9J1|MYCA_BACIU"  "sp|O75592|MYCB2_HUMAN" "sp|Q80ZA4|PKHL1_MOUSE" "sp|Q01499|POLG_BVDVS"
[13] "sp|P19712|POLG_CSFVA"  "sp|P21530|POLG_CSFVB"  "sp|P94459|PPSD_BACSU"  "sp|Q8WN22|PRKDC_CANFA" "sp|P0C6T4|R1A_BCHK4"   "sp|P0C6T8|R1A_CVBEN"
[19] "sp|P0C6T9|R1A_CVBLU"   "sp|P0C6U0|R1A_CVBM"    "sp|P0C6U1|R1A_CVBQ"    "sp|P0C6U3|R1A_CVHN1"   "sp|P0C6U4|R1A_CVHN2"   "sp|P0C6U5|R1A_CVHN5"
[25] "sp|P0C6U8|R1A_CVHSA"   "sp|Q80TF6|STAR9_MOUSE" "sp|Q5THJ4|VP13D_HUMAN" "sp|O88799|ZAN_MOUSE"

[[5]]
 [1] "sp|Q19319|CADH4_CAEEL" "sp|P0C6F1|DYH2_MOUSE"  "sp|Q63170|DYH7_RAT"    "sp|Q8IBG1|DYHC1_PLAF7" "sp|Q9C1M7|DYHC_ASHGO"  "sp|P34036|DYHC_DICDI"
 [7] "sp|Q8NWQ6|EBH_STAAW"   "sp|Q5HPA2|EBH_STAEQ"   "sp|Q8CP76|EBH_STAES"   "sp|P33450|FAT_DROME"   "sp|Q15751|HERC1_HUMAN" "sp|Q9VR91|HERC2_DROME"
[13] "sp|A2AJ76|HMCN2_MOUSE" "sp|Q18DN4|HMU_HALWD"   "sp|Q9JI18|LRP1B_MOUSE" "sp|P98157|LRP1_CHICK"  "sp|Q07954|LRP1_HUMAN"  "sp|Q91ZX7|LRP1_MOUSE"
[19] "sp|Q04833|LRP_CAEEL"   "sp|Q6TQR6|L_CCHFI"     "sp|Q12019|MDN1_YEAST"  "sp|Q8QGX4|PRKDC_CHICK" "sp|P97313|PRKDC_MOUSE" "sp|P0C6F3|R1A_BEV"
[25] "sp|P0C6U7|R1A_CVHOC"   "sp|P0C6U9|R1A_CVM2"    "sp|Q9P2P6|STAR9_HUMAN"
```

Comparativa

MODELO	5	10	50	100	500	1000	5000	10000	50000
Parallel.Data.Table	2.23	3.331	17.478	30.312	146.370	381.348	2555.176	6736.128	31940.7
Data.Table	2.448	15.162	25.104	123.121	232.023	1133.982	23009.212		
Plyr	13.43	21.046	110.6	202.194	983.3782	2612.60	20480.52		
Plyr + Doparallel	3.8	6.599	21.805	37.390	166.271	420.072	2694.451	7850.012	26675.06
RHadoop	120.352	115.142	117.367	95.604	125.883	196.19	661.689	1029.46	6171.976

Comparativa



Conclusión

- Cuantificar el alineamiento global de toda una base de datos de cadenas con una dada en busca de homología se ha conseguido.
- El tiempo de ejecución de este proceso ha sido reducido tanto en el modo local como en un clúster de máquinas.
- Comprobando la eficacia de Hadoop sobre una red de máquinas instauradas en la nube.
- Trabajo Futuro

GRACIAS