



Informatica

Principi di programmazione ed analisi dati in linguaggio
Scienze Biologiche

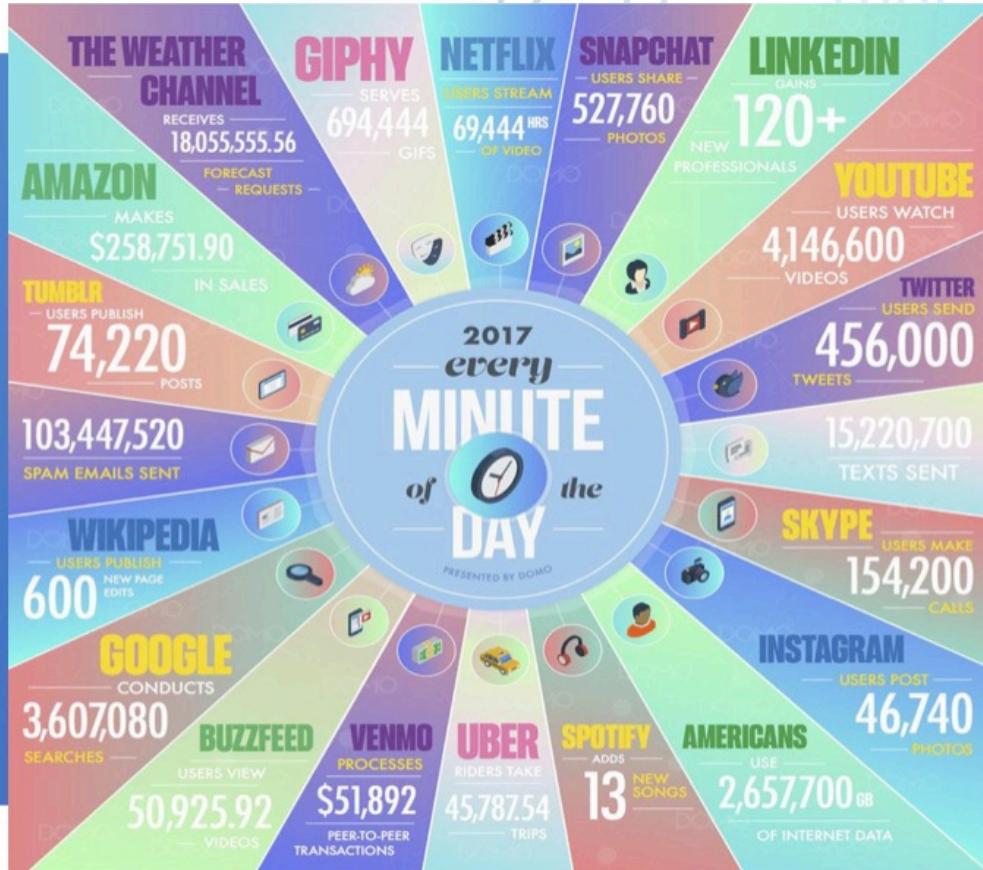


Dr. Bruno Bellisario, PhD

dens <- density(da dx <- dens\$x dy <- density infographic TRUE)

Big DATA

Perché parliamo di Big Data?...un minuto di Internet



Ogni minuto in Internet:

- Oltre di 4 milioni di filmati visti
- Oltre 3,5 milioni di richieste sui motori di ricerca
- 15 milioni di messaggi
- 103 milioni di email spam

Big DATA

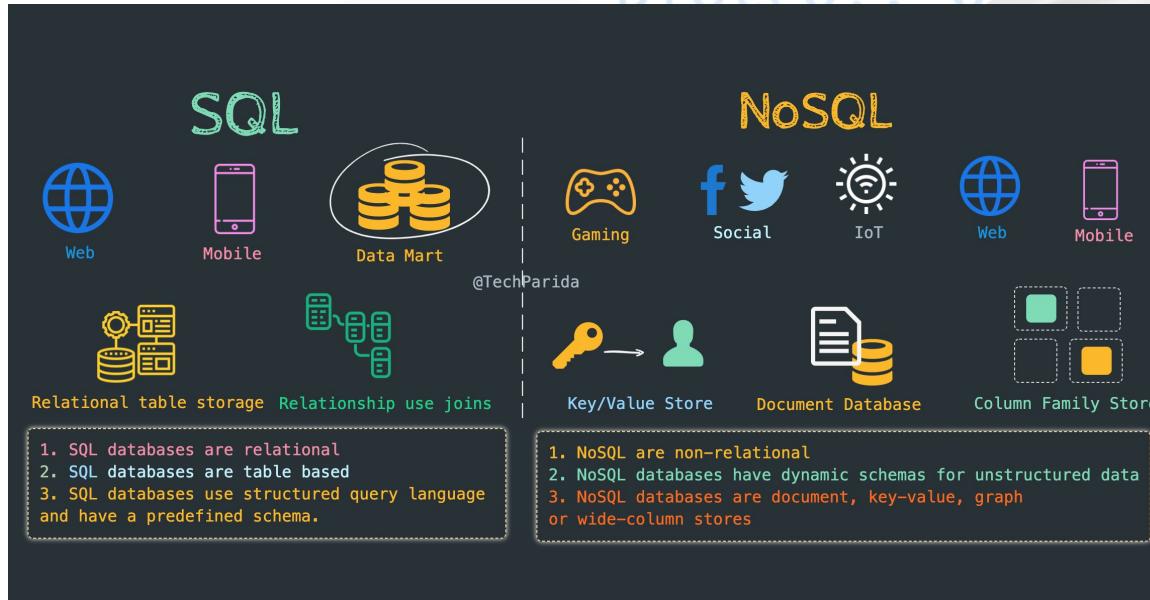
Definizione & Scenario

Il termine Big Data è usato per indicare un insieme di nuove tecnologie atte ad **acquisire, elaborare, mantenere e visualizzare** dati che altrimenti sarebbero intrattabili con le odierne tecnologie. Ciò è reso possibile grazie all'adozione di storage distribuiti, database NOSQL e in generale grazie alle tecniche di calcolo distribuito.



Big DATA

Definizione & Scenario



SQL è un metodo vecchio di decenni per accedere ai database relazionali.

Poiché i dati non strutturati, le quantità di archiviazione, la potenza di elaborazione e i tipi di analisi sono cambiati nel corso degli anni, si sono sviluppate diverse tecnologie di storage che si adattano meglio ai nuovi tipi di casi d'uso.

Questi database sono comunemente chiamati **NoSQL**.

SQL e NoSQL differiscono in quanto sono relazionali (SQL) o non relazionali (NoSQL), se i loro schemi sono predefiniti o dinamici, come scalano, il tipo di dati che includono e se sono più adatti per transazioni multi-riga o dati non strutturati.

Big DATA

Definizione & Scenario – le 5V dei Big Data

Volume

Il volume si riferisce alla quantità inimmaginabile di informazioni generate ogni secondo da social media, telefoni cellulari, automobili, carte di credito, immagini, video e quant'altro.

VOLUME

Huge amount of data

Varietà

I Big Data vengono generati in più varietà. Rispetto ai dati tradizionali come numeri di telefono e indirizzi, l'ultima tendenza dei dati è sotto forma di foto, video e audio e molti altri, rendendo circa l'80% dei dati completamente non strutturati.

Veridicità

La veridicità significa fondamentalmente il grado di affidabilità che i dati hanno da offrire. Poiché la maggior parte dei dati non è strutturata e irrilevante, i Big Data devono trovare un modo alternativo per filtrarli o tradurli in quanto i dati sono cruciali per gli sviluppi aziendali.

VERACITY

Inconsistencies and uncertainty in data

VELOCITY

High speed of accumulation of data

VARIETY

Different formats of data from various sources

Valore

Non è solo la quantità di dati che archiviamo o elaboriamo. In realtà è la quantità di dati preziosi, affidabili e degni di fiducia che devono essere archiviati, elaborati, analizzati per trovare approfondimenti.

Velocità

L'aspetto principale di Big Data è fornire dati su richiesta e a un ritmo più veloce.



Big DATA

Usi pratici

Tariffazione dinamica nelle assicurazioni auto.

In ambito assicurativo si è oramai affermato il concetto di Connected Car, in cui l'auto è costantemente connessa alla rete e invia informazioni relative alle condizioni di guida, ma anche allo stato del mezzo e della strada.

L'auto è geolocalizzata tramite segnale GPS e ne viene tracciato nel tempo il percorso.

Queste informazioni possono essere utilizzate a vari scopi, dalla manutenzione predittiva all'individuazione di anomalie prima che queste creino problemi, alla digitalizzazione delle condizioni delle infrastrutture (ad esempio ricostruendo un modello virtuale delle strade sulla base dei dati degli ammortizzatori).



Big DATA

Usi pratici

Monitoraggio della Qualità del Prodotto Industriale

La chiave per utilizzare le piattaforme di Big Data Analytics per aumentare l'efficienza operativa è usarle per sbloccare le informazioni presenti nei dati di log, sensori, macchine, server, infrastrutture di rete di telecomunicazione ed energetica.

Queste informazioni includono informazioni su trend, modelli e valori anomali che possono migliorare le decisioni, migliorare le prestazioni operative, risparmiare sui costi operativi e abilitare nuovi servizi.



Big DATA

Usi pratici

Protezione del Territorio

Immagini satellitari e la presenza sul territorio di sensori di vario tipo, permettono di raccogliere grandi quantità di informazioni geolocalizzate che mappano il territorio in una rappresentazione digitale dello stesso.

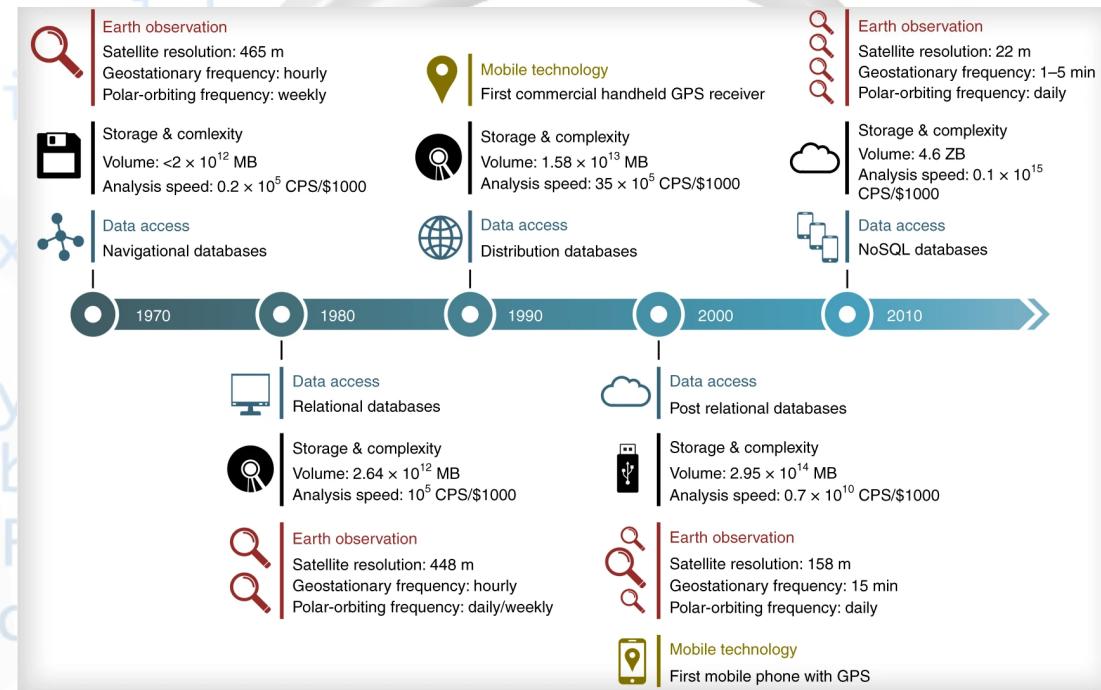
Mettendo in relazione questi dati e le informazioni sulla rete di comunicazioni, è possibile definire mappe di rischio e monitorare in tempo reale il territorio, ottimizzando di conseguenza dislocamento e percorsi dei mezzi di soccorso in caso di incendi, frane, allagamenti o altre emergenze.



dens <- density(da dx <- dens\$x Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

PER COMPRENDERE GLI IMPATTI DEGLI STRESSORI AMBIENTALI A LIVELLO GLOBALE (AD ESEMPIO I CAMBIAMENTI CLIMATICI, FLOTTE DA PESCA INTERNAZIONALI), ABBIAMO BISOGNO DI GRANDI DATASET GLOBALI



Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Cosa intendiamo, praticamente, per Big Data in campo ambientale?

Molti dati!

Una quantità **talmente elevata** per cui abbiamo bisogno di nuovi modi per elaborarli.

Sicuramente non a mano, probabilmente neanche con Excel, forse nemmeno sul tuo computer locale.



Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Come si ottengono i Big Data?

1. Combinando set di dati ridotti su larga scala per creare set di dati più dettagliati

Svantaggi: non sono necessariamente geograficamente e temporalmente coincidenti, rendendo necessario uno sforzo aggiuntivo per il loro confronto.

Vantaggi: altamente dettagliato



Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Come si ottengono i Big Data?

2. Dati da telerilevamento (ad es. immagini satellitari e droni)

Svantaggi: è possibile misurare solo determinate variabili e spesso non sono nemmeno dirette

Vantaggi: spesso si estendono su una grande quantità di spazio e tempo

RStoolbox: Tools for Remote Sensing Data Analysis in R

The Package

RStoolbox is a R package providing a wide range of tools for your every-day remote sensing processing needs. The available toolset covers many aspects from data import, pre-processing, data analysis, image classification and graphical display. RStoolbox builds upon the raster package, which makes it suitable for processing large data-sets even on smaller workstations. Moreover in most parts decent support for parallel processing is implemented. The package is published under GPL3.



Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Come si ottengono i Big Data?

3. Combinando misurazioni da sensori effettuate su area vasta (ad esempio galleggianti nell'oceano che misurano caratteristiche oceanografiche, stazioni meteorologiche, ecc.)

Svantaggi: alcuni dei dati necessari possono essere privati o scarsi perché gli strumenti sono costosi

Vantaggi: valori generalmente piuttosto standard/comparabili

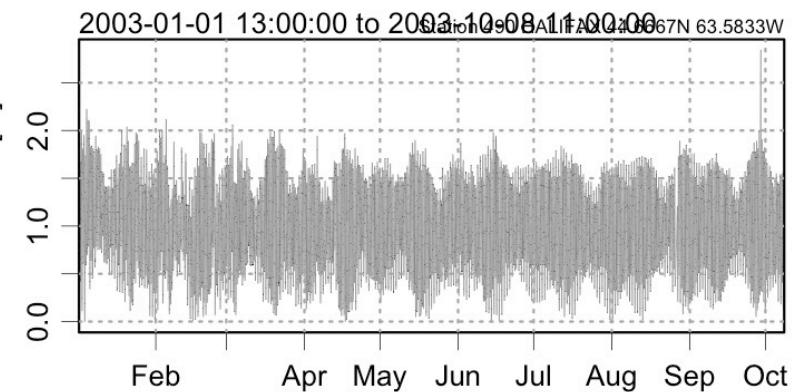
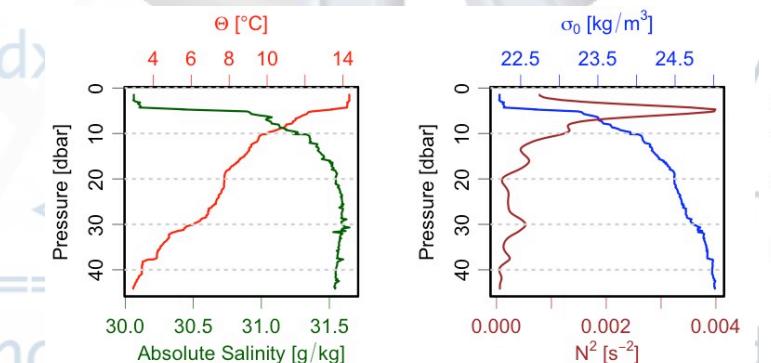
oce: an R package for Oceanographic Analysis

Dan E. Kelley^{1¶}, Clark Richards², and Chantelle Layton³

¹ Dan E. Kelley, Professor, Dalhousie University ² Clark Richards, Research Scientist, Bedford Institute of Oceanography, Department of Fisheries and Oceans, Canada; also Adjunct Professor, Dalhousie University ³ Chantelle Layton, Physical Scientist, Bedford Institute of Oceanography, Department of Fisheries and Oceans, Canada [¶] Corresponding author

Statement of Need

Oceanographic field experiments often employ a suite of instrument types, each reporting data in a different format. Many of these formats are complex and difficult to decode. Manufacturers usually provide software for accessing data produced by their instruments, but it is usually proprietary and closed-source, making it difficult for researchers to analyse their data in novel ways or to combine data from multiple instruments. The oce package (Kelley, Richards, & Layton, 2021) addresses such issues in the R language¹ with functions that handle dozens of data formats. It also has facilities for the specialized calculations and data displays that are particular to oceanography. Since oce is written in the R language (Ihaka & Gentleman, 1996; R Core Team, 2021), it forms a link to a vast array of general tools that oceanographers use in their work (Kelley, 2018).



Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Come si ottengono i Big Data?

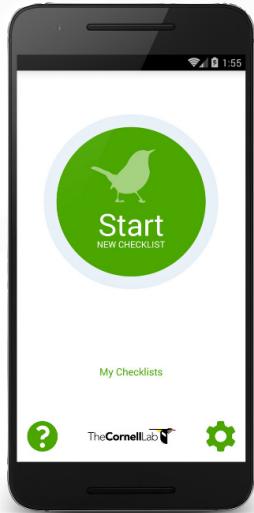
4. Programmi di *citizen science*: **eBird**, **Nature's Notebook**, **iNaturalist**

Grandi iniziative finanziate dai governi: **NEON** (National Ecological Observatory Network), **LTER** (ricerca ecologica a lungo termine)

Svantaggi: estremamente costoso, difficile da coordinare

Vantaggi: generalmente standardizzato (sebbene esistano alcune irregolarità)

eBird

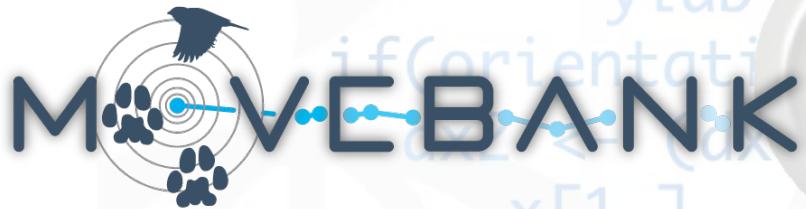


NATIONAL SCIENCE FOUNDATION
LTER NETWORK
LONG TERM ECOLOGICAL RESEARCH

Big DATA

Opportunità per i big data nella conservazione e nella sostenibilità

Alcuni esempi.



ecological networks database



Global Biodiversity
Information Facility



Big DATA

Come gestiamo i big data?

Computer !!

Software specifici basati su GUI (Excel/JMP/Stata/ecc.) o, meglio ancora, basato su script (R, Python, MATLAB)

Perché?

Perché la maggior parte della scienza non può essere ripetuta

1. Bisogna essere **trasparenti** al 100% in quello che si sta facendo ed essere in grado di controllare impostazioni, filtri, ecc.
2. Gli script forniscono una **copia digitale** di ciò che hai fatto con note per te e per gli altri in modo da ricordare perché hai fatto quello che hai fatto.
3. I tuoi passaggi sono scritti e chiunque può eseguirli riga per riga per vedere cosa succede
4. Le maggiori riviste stanno iniziando a richiedere la pubblicazione sia del codice che dei dati originali per la responsabilità.
5. **Non manipolare i dati originali:** un passo sbagliato e potresti rovinare l'intera catena di analisi!

Un esempio di Big DATA in biologia

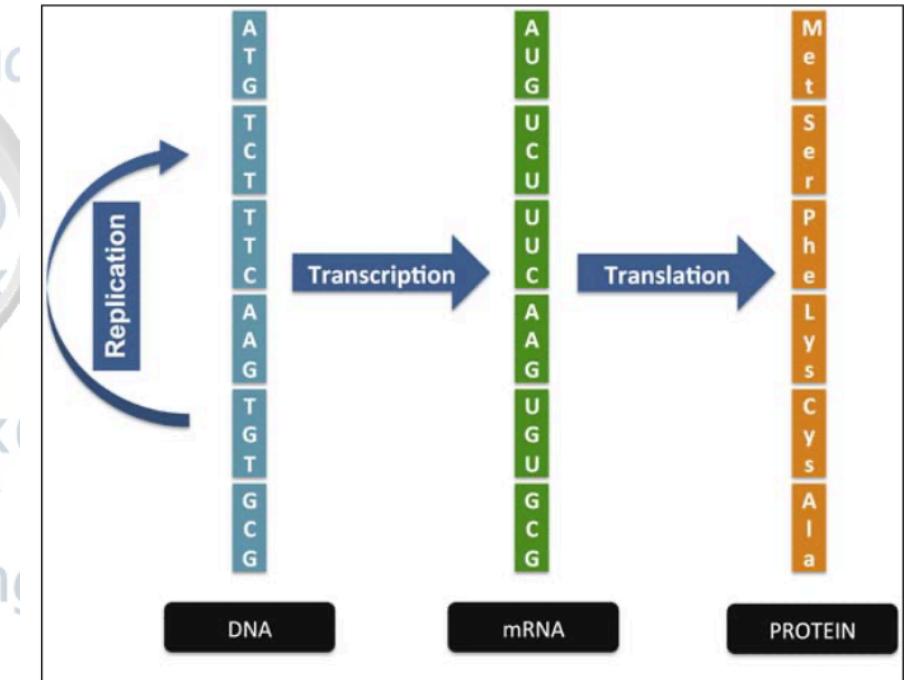
Dati molecolari

Gli acidi nucleici rappresentano geni, RNA e così via, mentre le proteine sono considerate i mattoni della vita.

Queste biomolecole rappresentano il contenuto informativo di un sistema vivente in termini di sequenze di caratteri.

E' la sequenza del DNA che determina la sequenza delle proteine. Ciò rende l'analisi delle sequenze (nucleotidiche e proteiche) importante per varie applicazioni che vanno dal confronto delle biomolecole per lo studio dell'evoluzione, alla mutazione e all'identificazione di siti interessanti nelle biomolecole, e così via.

L'enorme crescita dei dati delle sequenze ha aperto la strada all'evoluzione della **bioinformatica**.



Un esempio di Big DATA in biologia

Dati molecolari

La bioinformatica è un campo di studio interdisciplinare che combina il campo della biologia con l'informatica per la comprensione dei dati biologici.

La bioinformatica ha notevolmente accelerato il processo di scoperta ed è utilizzata sia in ambito accademico che nell'industria per comprendere i dati di genomica e proteomica, determinare la struttura delle proteine, le strutture dell'RNA, analizzare i dati di sequenza, l'identificazione dei geni, solo per citare alcune applicazioni.

Nel corso degli anni è diventato un campo disciplinare complesso che richiede un'istruzione e una formazione professionale specifica.



```
dy <- dens$y  
if(add == TRUE)  
  v1 <- 1  
  v2 <- 2  
  v3 <- 3  
  v4 <- 4  
  v5 <- 5  
  v6 <- 6  
  v7 <- 7  
  v8 <- 8  
  v9 <- 9  
  v10 <- 10  
  v11 <- 11  
  v12 <- 12  
  v13 <- 13  
  v14 <- 14  
  v15 <- 15  
  v16 <- 16  
  v17 <- 17  
  v18 <- 18  
  v19 <- 19  
  v20 <- 20  
  v21 <- 21  
  v22 <- 22  
  v23 <- 23  
  v24 <- 24  
  v25 <- 25  
  v26 <- 26  
  v27 <- 27  
  v28 <- 28  
  v29 <- 29  
  v30 <- 30  
  v31 <- 31  
  v32 <- 32  
  v33 <- 33  
  v34 <- 34  
  v35 <- 35  
  v36 <- 36  
  v37 <- 37  
  v38 <- 38  
  v39 <- 39  
  v40 <- 40  
  v41 <- 41  
  v42 <- 42  
  v43 <- 43  
  v44 <- 44  
  v45 <- 45  
  v46 <- 46  
  v47 <- 47  
  v48 <- 48  
  v49 <- 49  
  v50 <- 50  
  v51 <- 51  
  v52 <- 52  
  v53 <- 53  
  v54 <- 54  
  v55 <- 55  
  v56 <- 56  
  v57 <- 57  
  v58 <- 58  
  v59 <- 59  
  v60 <- 60  
  v61 <- 61  
  v62 <- 62  
  v63 <- 63  
  v64 <- 64  
  v65 <- 65  
  v66 <- 66  
  v67 <- 67  
  v68 <- 68  
  v69 <- 69  
  v70 <- 70  
  v71 <- 71  
  v72 <- 72  
  v73 <- 73  
  v74 <- 74  
  v75 <- 75  
  v76 <- 76  
  v77 <- 77  
  v78 <- 78  
  v79 <- 79  
  v80 <- 80  
  v81 <- 81  
  v82 <- 82  
  v83 <- 83  
  v84 <- 84  
  v85 <- 85  
  v86 <- 86  
  v87 <- 87  
  v88 <- 88  
  v89 <- 89  
  v90 <- 90  
  v91 <- 91  
  v92 <- 92  
  v93 <- 93  
  v94 <- 94  
  v95 <- 95  
  v96 <- 96  
  v97 <- 97  
  v98 <- 98  
  v99 <- 99  
  v100 <- 100  
  v101 <- 101  
  v102 <- 102  
  v103 <- 103  
  v104 <- 104  
  v105 <- 105  
  v106 <- 106  
  v107 <- 107  
  v108 <- 108  
  v109 <- 109  
  v110 <- 110  
  v111 <- 111  
  v112 <- 112  
  v113 <- 113  
  v114 <- 114  
  v115 <- 115  
  v116 <- 116  
  v117 <- 117  
  v118 <- 118  
  v119 <- 119  
  v120 <- 120  
  v121 <- 121  
  v122 <- 122  
  v123 <- 123  
  v124 <- 124  
  v125 <- 125  
  v126 <- 126  
  v127 <- 127  
  v128 <- 128  
  v129 <- 129  
  v130 <- 130  
  v131 <- 131  
  v132 <- 132  
  v133 <- 133  
  v134 <- 134  
  v135 <- 135  
  v136 <- 136  
  v137 <- 137  
  v138 <- 138  
  v139 <- 139  
  v140 <- 140  
  v141 <- 141  
  v142 <- 142  
  v143 <- 143  
  v144 <- 144  
  v145 <- 145  
  v146 <- 146  
  v147 <- 147  
  v148 <- 148  
  v149 <- 149  
  v150 <- 150  
  v151 <- 151  
  v152 <- 152  
  v153 <- 153  
  v154 <- 154  
  v155 <- 155  
  v156 <- 156  
  v157 <- 157  
  v158 <- 158  
  v159 <- 159  
  v160 <- 160  
  v161 <- 161  
  v162 <- 162  
  v163 <- 163  
  v164 <- 164  
  v165 <- 165  
  v166 <- 166  
  v167 <- 167  
  v168 <- 168  
  v169 <- 169  
  v170 <- 170  
  v171 <- 171  
  v172 <- 172  
  v173 <- 173  
  v174 <- 174  
  v175 <- 175  
  v176 <- 176  
  v177 <- 177  
  v178 <- 178  
  v179 <- 179  
  v180 <- 180  
  v181 <- 181  
  v182 <- 182  
  v183 <- 183  
  v184 <- 184  
  v185 <- 185  
  v186 <- 186  
  v187 <- 187  
  v188 <- 188  
  v189 <- 189  
  v190 <- 190  
  v191 <- 191  
  v192 <- 192  
  v193 <- 193  
  v194 <- 194  
  v195 <- 195  
  v196 <- 196  
  v197 <- 197  
  v198 <- 198  
  v199 <- 199  
  v200 <- 200  
  v201 <- 201  
  v202 <- 202  
  v203 <- 203  
  v204 <- 204  
  v205 <- 205  
  v206 <- 206  
  v207 <- 207  
  v208 <- 208  
  v209 <- 209  
  v210 <- 210  
  v211 <- 211  
  v212 <- 212  
  v213 <- 213  
  v214 <- 214  
  v215 <- 215  
  v216 <- 216  
  v217 <- 217  
  v218 <- 218  
  v219 <- 219  
  v220 <- 220  
  v221 <- 221  
  v222 <- 222  
  v223 <- 223  
  v224 <- 224  
  v225 <- 225  
  v226 <- 226  
  v227 <- 227  
  v228 <- 228  
  v229 <- 229  
  v230 <- 230  
  v231 <- 231  
  v232 <- 232  
  v233 <- 233  
  v234 <- 234  
  v235 <- 235  
  v236 <- 236  
  v237 <- 237  
  v238 <- 238  
  v239 <- 239  
  v240 <- 240  
  v241 <- 241  
  v242 <- 242  
  v243 <- 243  
  v244 <- 244  
  v245 <- 245  
  v246 <- 246  
  v247 <- 247  
  v248 <- 248  
  v249 <- 249  
  v250 <- 250  
  v251 <- 251  
  v252 <- 252  
  v253 <- 253  
  v254 <- 254  
  v255 <- 255  
  v256 <- 256  
  v257 <- 257  
  v258 <- 258  
  v259 <- 259  
  v260 <- 260  
  v261 <- 261  
  v262 <- 262  
  v263 <- 263  
  v264 <- 264  
  v265 <- 265  
  v266 <- 266  
  v267 <- 267  
  v268 <- 268  
  v269 <- 269  
  v270 <- 270  
  v271 <- 271  
  v272 <- 272  
  v273 <- 273  
  v274 <- 274  
  v275 <- 275  
  v276 <- 276  
  v277 <- 277  
  v278 <- 278  
  v279 <- 279  
  v280 <- 280  
  v281 <- 281  
  v282 <- 282  
  v283 <- 283  
  v284 <- 284  
  v285 <- 285  
  v286 <- 286  
  v287 <- 287  
  v288 <- 288  
  v289 <- 289  
  v290 <- 290  
  v291 <- 291  
  v292 <- 292  
  v293 <- 293  
  v294 <- 294  
  v295 <- 295  
  v296 <- 296  
  v297 <- 297  
  v298 <- 298  
  v299 <- 299  
  v300 <- 300  
  v301 <- 301  
  v302 <- 302  
  v303 <- 303  
  v304 <- 304  
  v305 <- 305  
  v306 <- 306  
  v307 <- 307  
  v308 <- 308  
  v309 <- 309  
  v310 <- 310  
  v311 <- 311  
  v312 <- 312  
  v313 <- 313  
  v314 <- 314  
  v315 <- 315  
  v316 <- 316  
  v317 <- 317  
  v318 <- 318  
  v319 <- 319  
  v320 <- 320  
  v321 <- 321  
  v322 <- 322  
  v323 <- 323  
  v324 <- 324  
  v325 <- 325  
  v326 <- 326  
  v327 <- 327  
  v328 <- 328  
  v329 <- 329  
  v330 <- 330  
  v331 <- 331  
  v332 <- 332  
  v333 <- 333  
  v334 <- 334  
  v335 <- 335  
  v336 <- 336  
  v337 <- 337  
  v338 <- 338  
  v339 <- 339  
  v340 <- 340  
  v341 <- 341  
  v342 <- 342  
  v343 <- 343  
  v344 <- 344  
  v345 <- 345  
  v346 <- 346  
  v347 <- 347  
  v348 <- 348  
  v349 <- 349  
  v350 <- 350  
  v351 <- 351  
  v352 <- 352  
  v353 <- 353  
  v354 <- 354  
  v355 <- 355  
  v356 <- 356  
  v357 <- 357  
  v358 <- 358  
  v359 <- 359  
  v360 <- 360  
  v361 <- 361  
  v362 <- 362  
  v363 <- 363  
  v364 <- 364  
  v365 <- 365  
  v366 <- 366  
  v367 <- 367  
  v368 <- 368  
  v369 <- 369  
  v370 <- 370  
  v371 <- 371  
  v372 <- 372  
  v373 <- 373  
  v374 <- 374  
  v375 <- 375  
  v376 <- 376  
  v377 <- 377  
  v378 <- 378  
  v379 <- 379  
  v380 <- 380  
  v381 <- 381  
  v382 <- 382  
  v383 <- 383  
  v384 <- 384  
  v385 <- 385  
  v386 <- 386  
  v387 <- 387  
  v388 <- 388  
  v389 <- 389  
  v390 <- 390  
  v391 <- 391  
  v392 <- 392  
  v393 <- 393  
  v394 <- 394  
  v395 <- 395  
  v396 <- 396  
  v397 <- 397  
  v398 <- 398  
  v399 <- 399  
  v400 <- 400  
  v401 <- 401  
  v402 <- 402  
  v403 <- 403  
  v404 <- 404  
  v405 <- 405  
  v406 <- 406  
  v407 <- 407  
  v408 <- 408  
  v409 <- 409  
  v410 <- 410  
  v411 <- 411  
  v412 <- 412  
  v413 <- 413  
  v414 <- 414  
  v415 <- 415  
  v416 <- 416  
  v417 <- 417  
  v418 <- 418  
  v419 <- 419  
  v420 <- 420  
  v421 <- 421  
  v422 <- 422  
  v423 <- 423  
  v424 <- 424  
  v425 <- 425  
  v426 <- 426  
  v427 <- 427  
  v428 <- 428  
  v429 <- 429  
  v430 <- 430  
  v431 <- 431  
  v432 <- 432  
  v433 <- 433  
  v434 <- 434  
  v435 <- 435  
  v436 <- 436  
  v437 <- 437  
  v438 <- 438  
  v439 <- 439  
  v440 <- 440  
  v441 <- 441  
  v442 <- 442  
  v443 <- 443  
  v444 <- 444  
  v445 <- 445  
  v446 <- 446  
  v447 <- 447  
  v448 <- 448  
  v449 <- 449  
  v450 <- 450  
  v451 <- 451  
  v452 <- 452  
  v453 <- 453  
  v454 <- 454  
  v455 <- 455  
  v456 <- 456  
  v457 <- 457  
  v458 <- 458  
  v459 <- 459  
  v460 <- 460  
  v461 <- 461  
  v462 <- 462  
  v463 <- 463  
  v464 <- 464  
  v465 <- 465  
  v466 <- 466  
  v467 <- 467  
  v468 <- 468  
  v469 <- 469  
  v470 <- 470  
  v471 <- 471  
  v472 <- 472  
  v473 <- 473  
  v474 <- 474  
  v475 <- 475  
  v476 <- 476  
  v477 <- 477  
  v478 <- 478  
  v479 <- 479  
  v480 <- 480  
  v481 <- 481  
  v482 <- 482  
  v483 <- 483  
  v484 <- 484  
  v485 <- 485  
  v486 <- 486  
  v487 <- 487  
  v488 <- 488  
  v489 <- 489  
  v490 <- 490  
  v491 <- 491  
  v492 <- 492  
  v493 <- 493  
  v494 <- 494  
  v495 <- 495  
  v496 <- 496  
  v497 <- 497  
  v498 <- 498  
  v499 <- 499  
  v500 <- 500  
  v501 <- 501  
  v502 <- 502  
  v503 <- 503  
  v504 <- 504  
  v505 <- 505  
  v506 <- 506  
  v507 <- 507  
  v508 <- 508  
  v509 <- 509  
  v510 <- 510  
  v511 <- 511  
  v512 <- 512  
  v513 <- 513  
  v514 <- 514  
  v515 <- 515  
  v516 <- 516  
  v517 <- 517  
  v518 <- 518  
  v519 <- 519  
  v520 <- 520  
  v521 <- 521  
  v522 <- 522  
  v523 <- 523  
  v524 <- 524  
  v525 <- 525  
  v526 <- 526  
  v527 <- 527  
  v528 <- 528  
  v529 <- 529  
  v530 <- 530  
  v531 <- 531  
  v532 <- 532  
  v533 <- 533  
  v534 <- 534  
  v535 <- 535  
  v536 <- 536  
  v537 <- 537  
  v538 <- 538  
  v539 <- 539  
  v540 <- 540  
  v541 <- 541  
  v542 <- 542  
  v543 <- 543  
  v544 <- 544  
  v545 <- 545  
  v546 <- 546  
  v547 <- 547  
  v548 <- 548  
  v549 <- 549  
  v550 <- 550  
  v551 <- 551  
  v552 <- 552  
  v553 <- 553  
  v554 <- 554  
  v555 <- 555  
  v556 <- 556  
  v557 <- 557  
  v558 <- 558  
  v559 <- 559  
  v560 <- 560  
  v561 <- 561  
  v562 <- 562  
  v563 <- 563  
  v564 <- 564  
  v565 <- 565  
  v566 <- 566  
  v567 <- 567  
  v568 <- 568  
  v569 <- 569  
  v570 <- 570  
  v571 <- 571  
  v572 <- 572  
  v573 <- 573  
  v574 <- 574  
  v575 <- 575  
  v576 <- 576  
  v577 <- 577  
  v578 <- 578  
  v579 <- 579  
  v580 <- 580  
  v581 <- 581  
  v582 <- 582  
  v583 <- 583  
  v584 <- 584  
  v585 <- 585  
  v586 <- 586  
  v587 <- 587  
  v588 <- 588  
  v589 <- 589  
  v590 <- 590  
  v591 <- 591  
  v592 <- 592  
  v593 <- 593  
  v594 <- 594  
  v595 <- 595  
  v596 <- 596  
  v597 <- 597  
  v598 <- 598  
  v599 <- 599  
  v600 <- 600  
  v601 <- 601  
  v602 <- 602  
  v603 <- 603  
  v604 <- 604  
  v605 <- 605  
  v606 <- 606  
  v607 <- 607  
  v608 <- 608  
  v609 <- 609  
  v610 <- 610  
  v611 <- 611  
  v612 <- 612  
  v613 <- 613  
  v614 <- 614  
  v615 <- 615  
  v616 <- 616  
  v617 <- 617  
  v618 <- 618  
  v619 <- 619  
  v620 <- 620  
  v621 <- 621  
  v622 <- 622  
  v623 <- 623  
  v624 <- 624  
  v625 <- 625  
  v626 <- 626  
  v627 <- 627  
  v628 <- 628  
  v629 <- 629  
  v630 <- 630  
  v631 <- 631  
  v632 <- 632  
  v633 <- 633  
  v634 <- 634  
  v635 <- 635  
  v636 <- 636  
  v637 <- 637  
  v638 <- 638  
  v639 <- 639  
  v640 <- 640  
  v641 <- 641  
  v642 <- 642  
  v643 <- 643  
  v644 <- 644  
  v645 <- 645  
  v646 <- 646  
  v647 <- 647  
  v648 <- 648  
  v649 <- 649  
  v650 <- 650  
  v651 <- 651  
  v652 <- 652  
  v653 <- 653  
  v654 <- 654  
  v655 <- 655  
  v656 <- 656  
  v657 <- 657  
  v658 <- 658  
  v659 <- 659  
  v660 <- 660  
  v661 <- 661  
  v662 <- 662  
  v663 <- 663  
  v664 <- 664  
  v665 <- 665  
  v666 <- 666  
  v667 <- 667  
  v668 <- 668  
  v669 <- 669  
  v670 <- 670  
  v671 <- 671  
  v672 <- 672  
  v673 <- 673  
  v674 <- 674  
  v675 <- 675  
  v676 <- 676  
  v677 <- 677  
  v678 <- 678  
  v679 <- 679  
  v680 <- 680  
  v681 <- 681  
  v682 <- 682  
  v683 <- 683  
  v684 <- 684  
  v685 <- 685  
  v686 <- 686  
  v687 <- 687  
  v688 <- 688  
  v689 <- 689  
  v690 <- 690  
  v691 <- 691  
  v692 <- 692  
  v693 <- 693  
  v694 <- 694  
  v695 <- 695  
  v696 <- 696  
  v697 <- 697  
  v698 <- 698  
  v699 <- 699  
  v700 <- 700  
  v701 <- 701  
  v702 <- 702  
  v703 <- 703  
  v704 <- 704  
  v705 <- 705  
  v706 <- 706  
  v707 <- 707  
  v708 <- 708  
  v709 <- 709  
  v710 <- 710  
  v711 <- 711  
  v712 <- 712  
  v713 <- 713  
  v714 <- 714  
  v715 <- 715  
  v716 <- 716  
  v717 <- 717  
  v718 <- 718  
  v719 <- 719  
  v720 <- 720  
  v721 <- 721  
  v722 <- 722  
  v723 <- 723  
  v724 <- 724  
  v725 <- 725  
  v726 <- 726  
  v727 <- 727  
  v728 <- 728  
  v729 <- 729  
  v730 <- 730  
  v731 <- 731  
  v732 <- 732  
  v733 <- 733  
  v734 <- 734  
  v735 <- 735  
  v736 <- 736  
  v737 <- 737  
  v738 <- 738  
  v739 <- 739  
  v740 <- 740  
  v741 <- 741  
  v742 <- 742  
  v743 <- 743  
  v744 <- 744  
  v745 <- 745  
  v746 <- 746  
  v747 <- 747  
  v748 <- 748  
  v749 <- 749  
  v750 <- 750  
  v751 <- 751  
  v752 <- 752  
  v753 <- 753  
  v754 <- 754  
  v755 <- 755  
  v756 <- 756  
  v757 <- 757  
  v758 <- 758  
  v759 <- 759  
  v760 <- 760  
  v761 <- 761  
  v762 <- 762  
  v763 <- 763  
  v764 <- 764  
  v765 <- 765  
  v766 <-
```

Un esempio di Big DATA in biologia

Dati molecolari

I pacchetti R per la bioinformatica: **Bioconductor** e **SeqinR**

Esistono molti pacchetti R per eseguire un'ampia varietà di analisi. Questi non vengono forniti con l'installazione standard di R, ma devono essere installati e caricati come "add-on", normalmente con il comando: `install.packages(...)`.

Bioconductor (<https://www.bioconductor.org>), un sito che contiene diversi pacchetti con molte funzioni R per l'analisi di set di dati biologici;

SeqinR è invece un pacchetto di funzioni R per recuperare ed analizzare sequenze di DNA e proteine da database specializzati.

Per utilizzare le funzioni del pacchetto SeqinR, dobbiamo prima installarlo con il comando: `install.packages("seqinr")`. Una volta installato, è possibile caricarlo per l'utilizzo digitando: `library(seqinr)`.

Un esempio di Big DATA in biologia

Dati molecolari

Il formato FASTA

Il formato **FASTA** è un formato semplice e ampiamente utilizzato per la memorizzazione di sequenze biologiche (DNA o proteine). È stato utilizzato per la prima volta dal programma **FASTA** per l'allineamento delle sequenze.

Inizia con una riga di descrizione che inizia con il carattere **>**, seguita da righe di sequenze.

Identifier

>NODE_1_length_1401924_cov_73.350607

Sequence

AAAGTCTCCTCACGCAAACCCTCTTGGTGGGTACCGAAGATAACCTGGCAAAGGA
GCACCTCTACCAGGGTCGTGAATGATCTGCTAACATATCCACGGATCTGTCCTTCACGT
TCGGAAGAGTCCAAGCTTAAGCTTGGCAGCTCCCTTCAGACGAGTGTGAGCGGTG
AAGATGGAACCAGCACCCTTCTTGAGCACGAATAACTCTACCCATGTTGTGTTAATGT
TTCTTGCTGTAAGAGGACTTGAAATTGTTATTGGTTTTGGGGAGTATGAGGG
TTCTTGTTGCGGGTTAACCTAGTGTGGTCACGTGCCTATTGGGCAAGCTGTGTG
AGGTATCATAAGGTGGTAGTTGAAAGGTACCTTATGGAAGACTTCGTTAGGAAGGTGTCT
GTATGATTAGAGTGGCGTAGGGTGAATGATTTAATTCTTCTTCG.....

Una sequenza inizia con un carattere maggiore di (">") seguito da una descrizione della sequenza (tutto su un'unica riga).

Le righe immediatamente successive alla riga di descrizione sono la rappresentazione della sequenza, con una lettera per amminoacido o acido nucleico, e in genere non superano gli 80 caratteri di lunghezza.

Un esempio di Big DATA in biologia

Dati molecolari

Il formato FASTA

Il formato **FASTA** ha, tipicamente, estensione del tipo:

Extension	Meaning	Notes
fasta, fa ^[9]	generic FASTA	Any generic fasta file. See below for other common FASTA file extensions
fna	FASTA nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	FASTA amino acid	Contains amino acid sequences. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

Un esempio di Big DATA in biologia

Dati molecolari

Il database di sequenze NCBI

Il *National Centre for Biotechnology Information (NCBI)* (www.ncbi.nlm.nih.gov) negli Stati Uniti gestisce un enorme database di tutti i dati raccolti sulle sequenze di DNA e proteine, il *NCBI Sequence Database*.

Ogni sequenza nel *NCBI Sequence Database* è memorizzata in un record separato, cui è assegnato un identificatore univoco che può essere utilizzato per fare riferimento a quella sequenza.

L'identificatore è noto come *accession* e consiste in una stringa di numeri e lettere. Ad esempio, il virus Dengue, che causa la febbre omonima, è presente in quattro tipi: DEN-1, DEN-2, DEN-3 e DEN-4. Gli accession NCBI per le sequenze di DNA dei virus DEN-1, DEN-2, DEN-3 e DEN-4 sono rispettivamente NC_001477, NC_001474, NC_001475 e NC_002640.

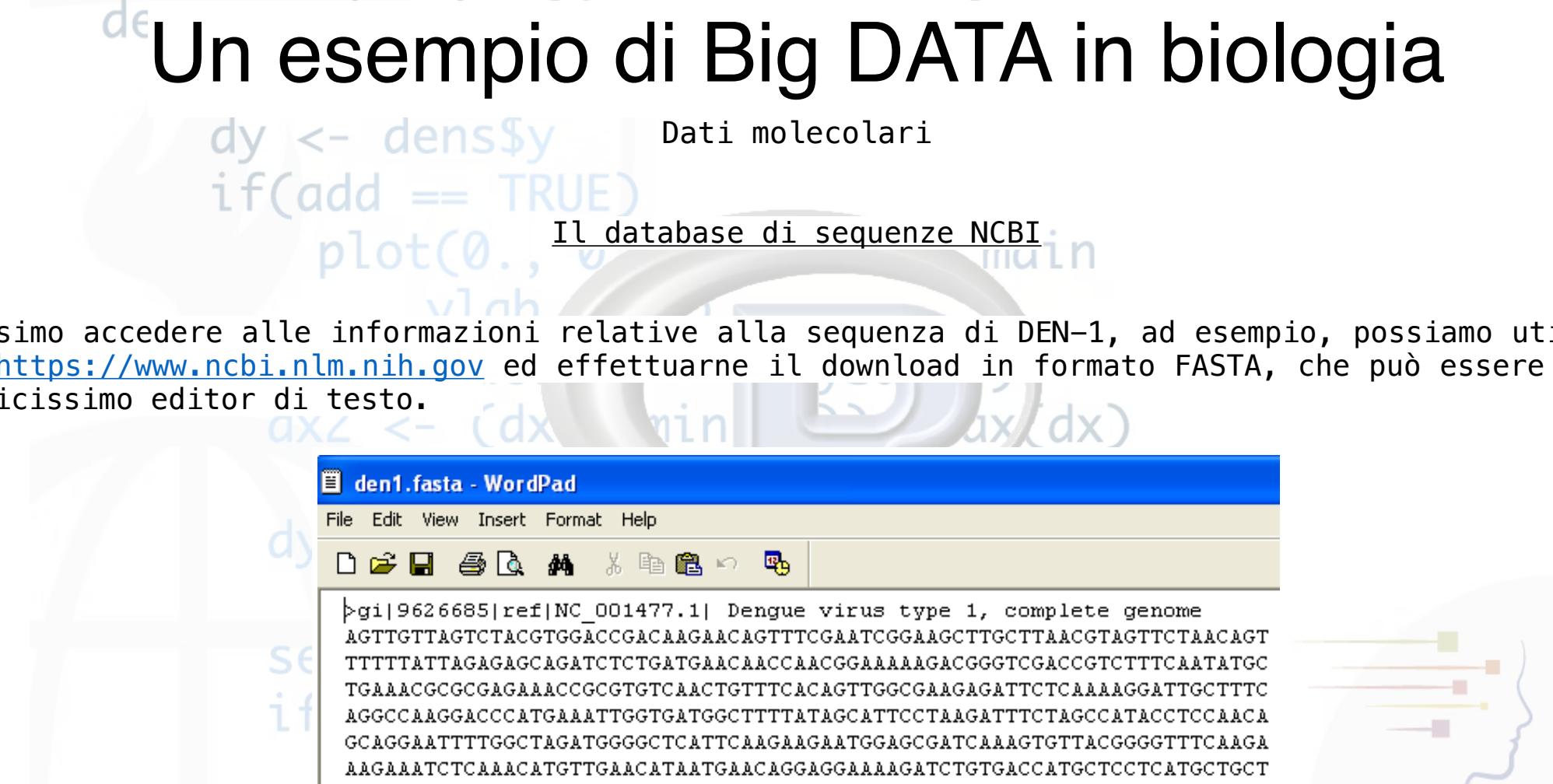


Un esempio di Big DATA in biologia

Dati molecolari

Il database di sequenze NCBI

Se volessimo accedere alle informazioni relative alla sequenza di DEN-1, ad esempio, possiamo utilizzare il portale <https://www.ncbi.nlm.nih.gov> ed effettuarne il download in formato FASTA, che può essere aperto con un semplicissimo editor di testo.



A screenshot of a Windows WordPad window titled "den1.fasta - WordPad". The window contains a single line of FASTA sequence data:

```
>gi|9626685|ref|NC_001477.1| Dengue virus type 1, complete genome
AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTCTGAATCGGAAGCTTGCTAACGTAAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTCATAATGC
TGAAACGCGCGAGAAACCGCGTGTCAAATGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCACATGAAATTGGTGTGGCTTTATAGCATTCTAACGATTCTAGCCATACCTCCAACA
GCAGGAATTGGCTAGATGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTACGGGGTTTCAGA
AAGAAATCTCAACATGTTGAACATAATGAAACAGGAGGAAAAGATCTGTGACCATGCTCTCATGCTGCT
GCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTGTTAACGACCTCTGCAGGTGTCAACATGTGCACCCATTGCAATGGATTGG
```

Un esempio di Big DATA in biologia

Dati molecolari

Recupero di sequenze genomiche tramite il pacchetto SeqinR

I dati di una sequenza possono essere recuperati sia visitando il sito web NCBI come visto in precedenza, sia accedendo al database all'interno di R tramite i comandi/funzioni del pacchetto **SeqinR**. Per utilizzarli, carichiamo prima la libreria con il seguente comando:

```
install.packages("seqinr")
library(seqinr)
```

Quindi, per selezionare la banca dati necessaria per ottenere la sequenza, possiamo controllare quelle disponibili con il comando

```
choosebank()
```

Supponiamo di voler utilizzare la banca dati genomica **genbank**

```
choosebank("genbank", timeout=20) #il parametro timeout serve qualora i tempi di risposta del server remoto siano lenti
```

Un esempio di Big DATA in biologia

Dati molecolari

Recupero di sequenze genomiche tramite il pacchetto SeqinR

Una volta aperta la banca dati remota, la possiamo interrogare con il seguente comando (questa fase richiede tempo per essere completata), che cerca le sequenze registrate per il gene BRCA1 della specie *Homo sapiens* (gene che codifica una fosfoproteina fondamentale per la stabilità genomica e come soppressore di tumori):

```
BRCA1 <- query("BRCA1", "SP=Homo sapiens AND K=BRCA1")
```

Le componenti dell'oggetto risultanti dall'interrogazione possono essere visualizzate con il comando:

```
attributes(BRCA1)
```

```
$names  
[1] "call"      "name"      "nelem"     "typelist"  "req"       "socket"  
$class  
[1] "qaw"
```



Un esempio di Big DATA in biologia

Dati molecolari

Recupero di sequenze genomiche tramite il pacchetto SeqinR

Creiamo una variabile `my_seqs` per leggere **tutte** le sequenze recuperate:

```
myseqs <- getSequence(BRCA1)
```

Se volessimo recuperare una sequenza specifica in base al suo *accession number* dovremmo digitare:

```
tmp <- query("tmp", "SP=Homo sapiens AND AC=U61268")#utilizziamo l'attributo AC nel comando di interrogazione
```

```
my_seq1 <- getSequence(tmp$req[[1]])#[[1]] recupera le informazioni relative al primo elemento della lista  
myseq1
```

```
[1] "t" "c" "g" "c" "t" "a" "g" "a" "a" "c" "c" "c" "g" "g" "g" "a" "g" "g" "c" "g" "g" "a" "g" "g" "t"  
[26] "t" "g" "c" "a" "g" "t" "g" "a" "g" "c" "c" "g" "a" "g" "a" "t" "c" "g" "c" "g" "c" "c" "a"  
"t" "t" ...  
[1301] "c" "t" "g" "g" "c" "c" "a" "a" "c" "a" "t" "g" "g" "t" "g" "a" "a" "a" "a" "c" "c" "c" "c" "t"  
[1326] "c" "t" "c" "c" "a" "c" "t" "a" "a" "a" "a" "a" "a" "t"
```

Un esempio di Big DATA in biologia

Dati molecolari

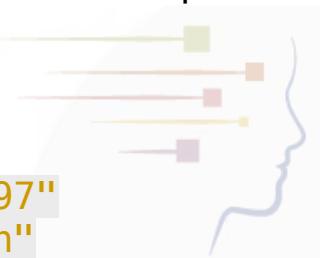
Recupero di sequenze genomiche tramite il pacchetto SeqinR

La banca dati selezionata viene interrogata tramite l'insieme dei termini di ricerca inseriti (K per parola chiave e SP per specie) e tutte le sequenze disponibili per gli attributi di ricerca indicati sono assegnate all'oggetto definito (nel nostro caso, "BRCA1"). Da questa lista possiamo filtrare i risultati rilevanti tramite i suoi attributi, come ad es. il nome o la posizione della sequenza (nel nostro caso abbiamo usato la posizione [[1]]).

Il comando `getSequence` recupera le sequenze dall'oggetto restituito dalla query.

Oltre all'attributo K (Keyword) per la query, ci sono altri possibili attributi che possono essere utilizzati (allo scopo basta visualizzare l'aiuto del comando "query"). È possibile recuperare ad esempio l'annotazione (gli altri attributi di una sequenza) con il comando `getAnnot`:

```
annots <- getAnnot(BRCA1$req[[1]])
annots
[1] "LOCUS HSU61268 1338 bp DNA linear PRI 16-JAN-1997"
[2] "DEFINITION Human breast and ovarian cancer susceptibility (BRCA1) gene, exon"
[3] "          2, partial flanking introns, and partial cds."
[4] "ACCESSION U61268"
```



Un esempio di Big DATA in biologia

Dati molecolari

Recupero di sequenze genomiche tramite il pacchetto SeqinR

Possiamo infine cercare gli identificatori delle sequenze (e visualizzarne il numero totale) come segue:

```
my_names <- getName(BRCA1)
length(my_names)
[1] 582
```

È sempre consigliabile chiudere la banca dati dopo l'interrogazione per evitare di avere più connessioni aperte. A tal fine è necessario utilizzare la seguente funzione di chiusura della banca dati (senza argomenti):

```
closebank()
```



Un esempio di Big DATA in biologia

Dati molecolari

Registrare i dati di una sequenza in un file FASTA

Dopo aver recuperato delle sequenze (e i relativi nomi), è possibile salvarle in un file in formato FASTA in R utilizzando la funzione `write.fasta()` del pacchetto **SeqinR**.

La funzione `write.fasta()` richiede il nome del file di output mediante l'argomento `file.out`. È inoltre necessario specificare la variabile `R` che contiene le sequenze (argomento `sequences`) e i nomi delle sequenze (argomento `names`).

Ad esempio, possiamo registrare le sequenze del gene BRCA1 recuperate nel paragrafo precedente (memorizzate nelle variabili `my_seqs` e `my_names`) nel file in formato FASTA **`brca1.fasta`** digitando:

```
write.fasta(sequences = my_seqs, names = my_names, file.out = "brca1.fasta")
```

```

seqbelow <- rep(y[1.], length(dx))
if(Fill == T)
    confshade(dx2, seqbelow, dy2

```

Un esempio di Big DATA in biologia

Dati molecolari

Leggere una sequenza da un file FASTA in R

Utilizzando il pacchetto **SeqinR** è possibile leggere facilmente una sequenza di DNA da un file FASTA in R. Per esempio, supponendo di voler leggere la sequenza di un particolare genoma da un file formato FASTA basta digitare il comando:

```
library(seqinr)  
my_fasta = read.fasta(file="NOME_FILE_FASTA")
```

Il comando sopra riportato legge il contenuto del file "NOME_FASTA_FILE" in un oggetto R di tipo lista chiamato `my_fasta`, che contiene le informazioni del file FASTA (il nome dato alla sequenza nel file e la sequenza di DNA stessa). Il primo elemento della lista contiene la sequenza DNA; quindi, possiamo memorizzare la sequenza di DNA in una variabile `my_fasta_seq` digitando:

```
my_fasta_seq = my_fasta[[1]]
```

Il comando sopra copia la sequenza di nucleotidi nel vettore `my_fasta_seq`.

Un esempio di Big DATA in biologia

dy <- dens\$y Dati molecolari

Dati molecolari

Brevi esem

Brevi esempi di utilizzo di file FASTA

Una volta recuperata una sequenza DNA, possiamo ottenere alcune semplici statistiche, come la lunghezza totale della sequenza in nucleotidi.

Recuperiamo la sequenza del genoma del virus DEN-1 e memorizziamola nella variabile vettore dengueseq. Per ottenere successivamente la lunghezza della sequenza del genoma, utilizziamo lo script successivo....

```
x[1.]  
dy2 <- (dx - min(dx)) / max(dx)  
y[1.]  
seqbelow <- rep(y[1.], length(dx))  
if(Fill == T)  
    confshade(dx2, seqbelow, dy2)
```

```
dens <- density(data, n = npts)

library(seqinr)      dx <- dens$x
choosebank("genbank", timeout=20)
my_fasta = read.fasta(file="sequence.fasta")#legge la sequenza del genoma del virus DEN-1 dal file in
# formato FASTA "sequence.fasta"
den1=my_fasta[[1]]#memorizziamo la sequenza di DNA per il virus DEN-1 in una variabile den1
#Una volta recuperata una sequenza DNA, possiamo ottenere alcune semplici statistiche, come la lunghezza
#totale della sequenza in nucleotidi (lunghezza sequenza del genoma)
length(den1)
#Una prima analisi ovvia di qualsiasi sequenza di DNA è quella di contare il numero di occorrenze dei
#quattro diversi nucleotidi ("A", "C", "G" e "T") nella sequenza. Questo può essere fatto usando la
#funzione table(). Per esempio, per trovare il numero di A, C, G e T nella sequenza del virus DEN-1
#(memorizzata in precedenza nella variabile den1), occorre digitare:
table(den1)
#da cui si deduce che la sequenza ha 127 nucleotidi A, 90 C, 105 G e 95 T.

#Possiamo anche conoscere la frequenza di ogni possibile coppia di nucleotidi nella sequenza utilizzando
#la funzione count(). Naturalmente è possibile farlo anche per le triple e così via scegliendo il giusto
#valore per l'argomento wordsize.
seqinr::count(den1, wordsize=2)#wordsize=2, coppie nucleotidiche
seqinr::count(den1, wordsize=3)#wordsize=3, triple nucleotidiche
closebank()
```

