```
Initialization complete (Lek-init.Rnw)

> system.time(load("../data/lek-ExpressionSet.RData"))
   user  system elapsed
   0.87    0.00    0.87
> ls(pattern = "^a")
[1] "allData"   "arrows.3d"
> eset <- allData

> nSamples <- dim(exprs(eset))[2]
```

2.9 Dostop do KEGG:

Dostop do KEGG v R omogoča paket **KEGGSOAP**, ki ga najdemo na Biocunductor strani.

S pomočjo funkcije bget lahko dosežemo dodatne informacije, ki jih prilepimo k opsu genov. Zaradi ??Pathways je zanimiv NCBI-GeneID, ki ga lahko povežemo s podpoljem Locus v polju Name GAL datoteke.

```
> eset <- allData[1:5, ]
> gName <- featureData(eset)$Name
> gName
[1] QC-3prime
[2] QC-5prime
[3] Locus:b0029|Gene:ispH|Prot:NP_414570_81_115
[4] Locus:b4267|Gene:idnD|Prot:NP_418688_253_287
[5] Locus:b3649|Gene:rpoZ|Prot:NP_418106_31_65
12041 Levels: <no oligo> ... QC-oligoB-AS-3
> grep("^Locus:", gName)
[1] 3 4 5
> Locus <- sub("^Locus:(b[0-9]*)[[:print:]]*", "\\1", gName)
> Locus[1:5]
[1] "QC-3prime" "QC-5prime" "b0029"     "b4267"     "b3649"
```

2.9.1 Dinamičen prevod posameznih kod

Takole bi lahko dinamično poizvedeli kakšne so posamezne kode za NCBI-GeneID:

```
> library(Hmisc)
> library(KEGGSOAP)
> n <- 5
> GeneID <- character(n)
> system.time(for (i in 1:n) {
+     x <- bget(paste("eco", Locus[i], sep = ":"))
+     if (length(x) > 0) {
+         first <- substring.location(x, "NCBI-GeneID: ")$last +
+             1
+         last <- substring.location(x, "\n", restrict = c(first,
+             9999))$first[1] - 1
+         GeneID[i] <- substring(x, first, last)
+     }
+ })
   user  system elapsed
   0.18    0.01    6.13
> cbind(Locus, GeneID)
     Locus       GeneID
[1,] "QC-3prime" ""
[2,] "QC-5prime" ""
[3,] "b0029"     "944777"
[4,] "b4267"     "944769"
[5,] "b3649"     "948160"
```

Za eno poizvedbo rabi malo več kot 1 sekundo.

2.9.2 Poizvedba za več genov naenkrat

Z eno poizvedbo lahko dobimo informacijo o več genih naenkrat, (nekje piše da največ 100). Za neobstoječe KEGG kode vrne prazen niz.

```
> eset <- allData[featureData(allData)$Status == "gene",
+     ]
> eset <- eset[1:100, ]
> gName <- featureData(eset)$Name
> Locus <- sub("^Locus:(b[0-9]*)[[:print:]]*", "\\1", gName)
> Locus[1:5]
[1] "b0029" "b4267" "b3649" "b0043" "b0083"
> stime <- system.time(seznam <- strsplit(bget(paste("eco",
+     Locus, sep = ":", collapse = " ")), "ENTRY"))
> stime
   user  system elapsed
   0.07    0.00    4.23
```

Dobili smo seznam kod v obliki `list`, preuredimo jo v matriko, pri čemer izpustimo redundantno vrstico.

```
> str(seznam)
List of 1
 $ : chr [1:100] "" "       b0029             CDS       E.coli\nNAME        is
> s <- as.matrix(seznam[[1]][-1], ncol = 1)
```

Za 100 kod je potreboval 4.23 sekund, tako da bi za vse gene potreboval približno 862.92 minut. Učinkoviteje, ampak še vedno veliko!

Vrne eno dodatno prazno vrstico na začetku. Pri nas imamo na mestu 95 neobstoječo kodo `b2999`, tako da je vrstni red rezultata drugačen od vrstnega reda vhodnih kod!!! Za neobstoječe ali neveljavne kode vrne prazen niz, zato vrstice ustrezajo zaporedju veljavnih KEGG kod. Lahko izločimo neveljavne (pri nas tovarniške kontrole) ali pa naredim preslikavo na podatke s pomočjo KEGG kode.

S funkcijo bget dobimo vso informacijo iz KEGG baze!

```
> cat(seznam[[1]][2])
        b0029                  CDS       E.coli
NAME        ispH, ECK0030, JW0027, lytB, yaaE
DEFINITION  1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase, 4Fe-4S
            protein (EC:1.17.1.2)
ORTHOLOGY   KO: K03527  4-hydroxy-3-methylbut-2-enyl diphosphate reductase
                [EC:1.17.1.2]
PATHWAY     PATH: eco00100  Biosynthesis of steroids
CLASS       Metabolism; Lipid Metabolism; Biosynthesis of steroids
            [PATH:eco00100]
POSITION    26277..27227
MOTIF       Pfam: LYTB
DBLINKS     Pasteur: ispH
            RegulonDB: B0029
            EcoGene: EG11081
            ECOCYC: EG11081
            NCBI-GI: 16128023
            NCBI-GeneID: 944777
            UniProt: P62623
CODON_USAGE       T                 C                 A                 G
          T   5   3   0   6   3   5   0   4   3   3   1   0   2   2   0   2
          C   0   2   0  15   0   0   0  12   2   4   0  12  14   9   0   0
          A  12   8   1   6   2   4   1   5   1  11  13   1   3   3   0   0
          G   5  10   5  17   3  11   6  10  11   8  22   7   7   9   1   5
AASEQ       316
            MQILLANPRGFCAGVDRAISIVENALAIYGAPIYVRHEVVHNRYVVDSLRERGAIFIEQI
            SEVPDGAILIFSAHGVSQAVRNEAKSRDLTVFDATCPLVTKVHMEVARASRRGEESILIG
            HAGHPEVEGTMGQYSNPEGGMYLVESPDDVWKLTVKNEEKLSFMTQTTLSVDDTSDVIDA
            LRKRFPKIVGPRKDDICYATTNRQEAVRALAEQAEVVLVVGSKNSSNSNRLAELAQRMGK
            RAFLIDDAKDIQEEWVKEVKCVGVTAGASAPDILVQNVVARLQQLGGGEAIPLEGREENI
            VFEVPKELRVDIREVD
NTSEQ        951
            atgcagatcctgttggccaacccgcgtggttttgtgccggggtagaccgcgctatcagc
            attgttgaaaacgcgctggccatttacggcgcaccgatatatgtccgtcacgaagtggta
            cataaccgctatgtggtcgatagcttgcgtgagcgtggggctatctttattgagcagatt
            agcgaagtaccggacggcgcgatcctgattttctccgcacacggtgtttctcaggcggta
            cgtaacgaagcaaaaagtcgcgatttgacggtgtttgatgccacctgtccgctggtgacc
            aaagtgcatatggaagtcgcccgcgccagtcgccgtggcgaagaatctattctcatcggt
            cacgccgggcacccggaagtggaagggacaatgggccagtacagtaacccggaaggggga
            atgtatctggtcgaatcgccggacgatgtgtggaaactgacggtcaaaaacgaagagaag
            ctctcctttatgacccagaccacgctgtcggtggatgacacgtctgatgtgatcgacgcg
            ctgcgtaaacgcttcccgaaaattgtcggtccgcgcaaagatgacatctgctacgccacg
            actaaccgtcaggaagcggtacgcgccctggcagaacaggcggaagttgtgttggtggtc
            ggttcgaaaaactcctccaactccaaccgtctggcggagctggcccagcgtatgggcaaa
            cgcgcgtttttgattgacgatgcgaaagacatccaggaagagtgggtgaaagaggttaaa
            tgcgtcggcgtgactgcgggcgcatcggctccggatattctggtgcagaatgtggtggca
            cgtttgcagcagctgggcggtggtgaagccattccgctggaaggccgtgaagaaaacatt
            gttttcgaagtgccgaaagagctgcgtgtcgatattcgtgaagtcgattaa
///
```

2.9.3 Izbira kod iz seznama: funkcija `getCode`

S funkcijo `getCode` lahko iz seznama dobimo NCBI-GeneID kode:

```
> getCode <- function(x, str1 = "NCBI-GeneID: ", str2 = "\n") {
+     first <- function(x, str = "NCBI-GeneID: ") substring.location(x,
+         str)$last + 1
+     last <- function(x, str = "\n", start = 1) (substring.location(x,
+         str, restrict = c(start, 9999))$first)[1] - 1
+     st <- first(x, str1)
+     en <- last(x, str2, start = st)
+     return(substring(x, st, en)[1])
+ }
```

Ob klicu funkcije `getCode` definiramo enoličen niz znakov pred kodo in niz, ki kodo zaključuje.

```
> GeneID <- apply(s, 1, getCode, "NCBI-GeneID: ", "\n")
> GeneID[1:5]
[1] "944777" "944769" "948160" "948958" "944803"
```

ali pa KEGG kode:

```
> KEGGId <- paste("b", apply(s, 1, getCode, "b", " "),
+     sep = "")
> KEGGId[1:5]
[1] "b0029" "b4267" "b3649" "b0043" "b0083"
```

in `ECOCYC` kode:

```
> ecocycId <- apply(s, 1, getCode, "ECOCYC: ")
> ecocycId[1:5]
[1] "EG11081" "G7893"   "EG10899" "EG11564" "EG11086"
```

Združena tabela

```
> print(cbind(KEGGId, GeneID, ecocycId)[1:5, ], quote = FALSE)
     KEGGId GeneID ecocycId
[1,] b0029  944777 EG11081
[2,] b4267  944769 G7893
[3,] b3649  948160 EG10899
[4,] b0043  948958 EG11564
[5,] b0083  944803 EG11086
```

Funkcijo `getCode` lahko uporabimo tudi na sestavljenih imenih genov:

```
> gName <- as.matrix(featureData(eset)$Name[1:5], ncol = 1)
> gName
```

```
     [,1]
[1,] "Locus:b0029|Gene:ispH|Prot:NP_414570_81_115"
[2,] "Locus:b4267|Gene:idnD|Prot:NP_418688_253_287"
[3,] "Locus:b3649|Gene:rpoZ|Prot:NP_418106_31_65"
[4,] "Locus:b0043|Gene:fixC|Prot:NP_414585_368_402"
[5,] "Locus:b0083|Gene:ftsL|Prot:NP_414625_21_56"
```

S primernimi oznakami začetkov in koncev lahko izluščimo kode:

```
> apply(gName, 1, getCode, "Locus:", "|")
[1] "b0029" "b4267" "b3649" "b0043" "b0083"

> apply(gName, 1, getCode, "Gene:", "|")
[1] "ispH" "idnD" "rpoZ" "fixC" "ftsL"
```

Včasih je treba biti nekoliko zvit :)

```
> paste("NP", apply(gName, 1, getCode, "Prot:NP_", "_"),
+     sep = "_")
[1] "NP_414570" "NP_418688" "NP_418106" "NP_414585" "NP_414625"
```

2.9.4 KEGG FTP

Na *FTP* strani `ftp://ftp.genome.jp/pub/kegg/genes/organisms/eco` pa najdemo šifrant za pretvorbo `LocusID` v `NCBI-GeneID` za E. coli `eco_ncbi-geneid.list`. To je seveda za prevedbo vseh genov precej ugodnejša varianta. Iz datoteke `eco_ncbi-geneid.list` bomo prepisali oznake v `featureData` za naše podatke.

2.9.5 Direktna povezava na KEGG

Seznam genov, ki imajo internetno povezavo na bazo KEGG Genes:

```
> ID <- Locus[1:5]
> KEGG <- "http://www.genome.jp/dbget-bin/www_bget?eco"
> links <- paste(KEGG, ID, sep = "+")
> refs <- paste("\\href{", links, "}{", ID, "}", sep = "",
+     collapse = "\\\\ \n")
> cat(refs, "\n")
```

b0029
b4267
b3649
b0043
b0083

Klik na ime pokaže podatke na spletni strani KEGG,
<ctrl>-klik pripravi stran s podatki v obliki PDF datoteke
<shift>-klik pa pripne PDF izpis spletne strani na konec tega dokumenta.