Exercise3-数据科学与数据分析-20230314

211294029 李梦麟

1.安装所需要的包： RMySQL

```
# install.packages("RMySQL", repos = "https://mirrors.ustc.edu.cn/CRAN/")
```

2.载入包

```
any(grepl("RMySQL",installed.packages())) # 查看RMySQL是否安装成功
```

```
## [1] TRUE
```

```
library(RMySQL)
```

3.连接 MySQL 数据库

```
# 建立R与MySQL的链接
mysqlconnection = dbConnect(MySQL(),
                            user='root', # MySQL数据库用户名
                            password='123456', # 对应用户的登录密码
                            dbname='paper', # 需要连接的数据库名
                            host='localhost', # 访问的数据库所在的IP
                            port=3306) # 访问的数据库所关联的端口号，一般为3306

# 查看链接信息
summary(mysqlconnection)
```

```
## <MySQLConnection:0,0>
##   User:   root
##   Host:   localhost
##   Dbname: paper
##   Connection type: localhost via TCP/IP
##
## Results:
```

```
# 查看该数据库内有哪些表
dbListTables(mysqlconnection)
```

```
## character(0)
```

4.使用 R 操作 MySQL 数据库的增删查改

```
# 在R环境中读取数据，为1000条论文文献数据
paper <- read.csv("paper.csv")

# 打印该数据的前6行
knitr::kable(head(paper))
```

id	st_title	st_abstract	st_year
s10084	neutralization-sensitive merozoite surface antigens of babesia bovis encoded by members of a polymorphic gene family.	neutralization-sensitive merozoite surface antigens of babesia bovis encoded by members of a polymorphic gene family.. Monospecific antibodies against native and recombinant versions of the major merozoite surface antigen (MSA-1) of Babesia bovis neutralize the infectivity of merozoites from Texas and Mexico strains in vitro. Sequence analysis shows that MSA-1 and a related, co-expressed 44 kDa merozoite surface protein (MSA-2) are encoded by members of a multigene family previously designated BabR. BabR genes, originally described in Australia strains of B. bovis, are notable because their marked polymorphism is apparently mediated by chromosomal rearrangements, but protein products of BabR genes have not previously been identified. The 3’ terminal 173 nucleotides of the MSA-1 gene, including 60 nucleotides of untranslated sequence, are highly similar to the 3’ terminal sequences of BabR 0.8 (84% identity) and MSA-2 (94% identity). Alignment of the predicted protein sequences demonstrates significant overall homology between MSA-1 and MSA-2, and between both proteins and the amino terminal BabR sequence. MSA-1 nucleic acid probes also hybridize weakly to genomic DNA from the Australia ‘L’ strain, even though this strain does not express merozoite	1992

id	st_title	st_abstract	st_year
		surface epitopes cross-reactive with MSA-1 or MSA-2. Hybridization of these same probes to genomic DNA from the cloned Mexico strain reveals a pattern of bands compatible with two copies each of MSA-1 and MSA-2. Proteins encoded by this B. bovis gene family have been designated variable merozoite surface antigens (VMSA). The extent and mechanism of VMSA polymorphism among strains will be important when evaluating the role these surface proteins have in the host-parasite interaction, including immunity to blood stages.	
s10093	signalling through the mhc class ii cytoplasmic domain is required for antigen presentation and induces b7 expression.	signalling through the mhc class ii cytoplasmic domain is required for antigen presentation and induces b7 expression.. Class II major histocompatibility complex (MHC) molecules function as antigen-presenting elements as well as signal transducers on B lymphocytes. We previously reported that a B lymphoma cell transfectant, 5C2, expressing genetically engineered I-Ak molecules with truncated cytoplasmic domains was severely impaired in both antigen presentation and in anti-Ia-induced intracytoplasmic signalling. These two functions could be restored by preculturing 5C2 cells with cyclic AMP analogues. Here we demonstrate that impaired signal transduction by truncated class II molecules results in a deficiency in induction of the newly defined B-cell accessory molecule B7 (ref. 8), which can be reversed by restoration of B7 expression. These data imply that contact of the T-cell antigen receptor with MHC/antigen ligand results in signal transmission through the class II cytoplasmic domain. This signal, which can be	1992

id	st_title	st_abstract	st_year
		mimicked by dibutyryl cAMP, induces expression of B7, resulting in effective antigen presentation. The fact that crosslinking of surface class II MHC also induces B7 expression on normal resting human B cells supports this contention.	
s10107	cloning and surface expression of pseudomonas aeruginosa o antigen in escherichia coli.	cloning and surface expression of pseudomonas aeruginosa o antigen in escherichia coli.. As a step toward developing recombinant oral vaccines, we have explored the feasibility of expression of O polysaccharide antigens from Pseudomonas aeruginosa by Escherichia coli. We cloned in E. coli HB101 a 26.2-kilobase DNA fragment from P. aeruginosa strain PA103 that specifies the production of the O polysaccharide of Fisher immunotype 2 (IT-2) strains. The recombinant organism incorporated the P. aeruginosa IT-2 O polysaccharide onto the core of the E. coli lipopolysaccharide (LPS). Transfer of the recombinant plasmid to three LPS-rough strains of P. aeruginosa resulted in synthesis of IT-2 O antigen, and two of these transconjugant strains also synthesized a second O polysaccharide, presumably representing expression of a repressed, or an incomplete, set of genes for an endogenous O polysaccharide. Rabbits injected with the purified recombinant LPS made antibody specific for P. aeruginosa IT-2 O side chains, as did mice fed the recombinant E. coli strain. Expression of P. aeruginosa O antigens by enteric bacteria makes it possible to study these recombinant strains as oral vaccines to prevent P. aeruginosa infections.	1992
s10198	murine b7 antigen	murine b7 antigen provides a	1992

id	st_title	st_abstract	st_year
	provides a sufficient costimulatory signal for antigen-specific and mhc-restricted t cell activation.	sufficient costimulatory signal for antigen-specific and mhc-restricted t cell activation.. We have previously shown that the murine B7 (mB7) molecule, when expressed in Chinese hamster ovary cells in stable fashion, can costimulate with anti-CD3 mAb or Con A to induce T cell activation. We have now derived, by gene transfection, Chinese hamster ovary cell lines that express the I-Ad molecule, either alone or in context with mB7. We have analyzed these transfectants for their capacity to present Ag to murine CD4+ T lymphocytes. I-Ad/mB7-double transfectants were able to stimulate mixed lymphocyte reactions and to present peptide Ag to specific T cells. Chinese hamster ovary cells that expressed only the I-Ad molecule were not able to stimulate T cell proliferation in these systems. Thus, the mB7 protein is a sufficient costimulatory molecule for the physiologic, Ag-dependent/MHC-restricted activation of murine CD4+ T cells. Stimulation of T cell bulk cultures resulted predominantly in the production of IL-2 and not of IL-4. The costimulatory activity of mB7 is not, however, restricted to the IL-2-secreting subset. We have identified one IL-4-secreting T cell clone, CDC35, which is responsive to mB7 triggering. Finally, we present experiments that suggest that mB7 and peptide/MHC complexes need to be expressed on the same cell for optimal induction of T cell activation.	
s10289	the blood group antigen-related glycoepitopes: key structural determinants	the blood group antigen-related glycoepitopes: key structural determinants in immunogenesis and aids pathogenesis.. This overview will	1992

id	st_title	st_abstract	st_year
	in immunogenesis and aids pathogenesis.	focus on the functional and pathophysiological aspects of blood group antigen (BGA)-related glycodeterminants with regard to immunogenesis and AIDS pathogenesis. It has been postulated that in a broad range of histogenetically different tissues and organs, BGA-related glycoepitopes are expressed on the cell surface at definite stages of cell differentiation. These glycoepitopes are expressed during embryogenesis, organogenesis, tissue repair, regeneration, remodelling and maturation when 'sorting-out' of one homotypic cell population from a heterotypic assemblage of cells occurs (1). In this event, the BGA-related glycoepitopes, if being expressed on the cell surface, play roles of key structural determinants in cell-cell recognition, association and aggregation. This mechanism will be discussed in relation to immunogenesis with regard to antigen presentation, self-non-self discrimination, and positive and negative selection during thymic education. It is postulated that the appearance of BGA-related glycoepitopes on the cell membrane is a consequence of the association of major histocompatibility complex antigens (MHC) and peptides, with the subsequent elimination of cells carrying a high density of BGA-related glycoepitopes on their surface. After human immunodeficiency virus (HIV) glycoproteins are glycosylated by host cell glycosyltransferases, the virus may use the BGA-related glycodeterminants as ligands and/or receptors for expansion to a spectrum of target cells during AIDS	

id	st_title	st_abstract	st_year
		development and generalization of the infection throughout the body. We will review the experimental evidence that supports the concept that HIV uses an alternative to the gp120/CD4 ligand/receptor system, and that the alternative mechanism is probably carbohydrate-mediated in nature.	
s10545	proteasome subunits encoded by the major histocompatibility complex are not essential for antigen presentation.	proteasome subunits encoded by the major histocompatibility complex are not essential for antigen presentation.. Major histocompatibility complex (MHC) class I molecules bind and deliver peptides derived from endogenously synthesized proteins to the cell surface for survey by cytotoxic T lymphocytes. It is believed that endogenous antigens are generally degraded in the cytosol, the resulting peptides being translocated into the endoplasmic reticulum where they bind to MHC class I molecules. Transporters containing an ATP-binding cassette encoded by the MHC class II region seem to be responsible for this transport. Genes coding for two subunits of the '20S' proteasome (a multicatalytic proteinase) have been found in the vicinity of the two transporter genes in the MHC class II region, indicating that the proteasome could be the unknown proteolytic entity in the cytosol involved in the generation of MHC class I-binding peptides. By introducing rat genes encoding the MHC-linked transporters into a human cell line lacking both transporter and proteasome subunit genes, we show here that the MHC-encoded proteasome subunit are not essential for stable MHC class I surface expression, or for processing and presentation of antigenic peptides from influenza virus and an	1992

id	st_title	st_abstract	st_year
		intracellular protein.	

```
# 操作一：在数据库中创建表
# 利用R的数据框构建数据库的结构
dbCreateTable(mysqlconnection,"paper",paper) # 新建一张名为paper的表
dbListTables(mysqlconnection)
```

```
## [1] "paper"
```

```
# 操作二：向数据表中插入数据
dbWriteTable(mysqlconnection,"paper",paper,row.names=FALSE,append=TRUE)
```

```
## [1] TRUE
```

```
# 注意：
# 如果报错：could not run statement: Loading local data is disabled; this must be enabled on both the client and server sides
# 需要调整MySQL全局参数，在MySQL中运行sql:set global local_infile=true;

# 操作三：查询,读取paper内的全部数据
paper_from_mysql = dbGetQuery(mysqlconnection, "select * from paper") # 在双引号内写SQL查询式
print(dim(paper_from_mysql))
```

```
## [1] 1000    4
```

```
# 注意：
# 如果报错：connection with pending rows, close resultSet before continuing
# 执行dbClearResult(dbListResults(连接名)[[1]])，清理查询结果，重新跑

# 操作四：删除表
dbRemoveTable(mysqlconnection,'paper')
```

```
## [1] TRUE
```

```
dbListTables(mysqlconnection)
```

```
## character(0)
```

5.关闭连接

```
dbDisconnect(mysqlconnection)
```

```
## [1] TRUE
```

6.练习

```
mysqlconnect = dbConnect(MySQL(),
                         user='root', # MySQL 数据库用户名
                         password='123456', # 对应用户的登录密码
                         dbname='paper', # 需要连接的数据库名
                         host='localhost', # 访问的数据库所在的 IP
                         port=3306) # 访问的数据库所关联的端口号，一般为 3306

dbCreateTable(mysqlconnect,"paper",paper)
dbWriteTable(mysqlconnect,"paper",paper,row.names=FALSE,append=TRUE)

## [1] TRUE

# 筛选 paper 数据中标题或摘要带有“viruse”关键词的论文数据，在数据库中使用筛选后的数据创建一张新表，新表的名称为 paper_viruse，并打印这张表的行数（样本数），以及前 6 行。

result <- dbGetQuery(mysqlconnect,"select* from paper where st_title like '%viruse%' or st_abstract like '%viruse%'" )
dbWriteTable(mysqlconnect,"paper_viruse",result,row.names=FALSE,append=TRUE)

## [1] TRUE

dbListTables(mysqlconnect)

## [1] "paper"        "paper_viruse"

paper.viruse <- dbReadTable(mysqlconnect,"paper_viruse")
nrow(paper.viruse)

## [1] 130

knitr::kable(head(paper.viruse))
```

id	st_title	st_abstract	st_year
s11120	comparison of a dengue-2 virus and its candidate vaccine derivative: sequence relationships with the flaviviruses and other viruses.	comparison of a dengue-2 virus and its candidate vaccine derivative: sequence relationships with the flaviviruses and other viruses.. A comparison of the sequence of the dengue-2 16681 virus with that of the candidate vaccine strain (16681-PDK53) derived from it identified 53 of the 10,723 nucleotides which differed between the strains. Nucleotide changes occurred in	1992

id	st_title	st_abstract	st_year
		genes coding for all virion and nonvirion proteins, and in the 5' and 3' untranslated regions. Twenty-seven of the nucleotide changes resulted in amino acid alterations. The greatest amino acid sequence differences in the virion proteins occurred in prM (2.20%; 2/91 amino acids) followed by the M protein (1.33%; 1/75 amino acids), the C protein (0.88%; 1/114 amino acid), and the E protein (0.61%; 3/495 amino acids). Differences in the amino acid sequence of nonvirion proteins ranged from 1.51% (6/398 amino acids) in NS4 to 0.33% (3/900 amino acids) in NS5. The encoded protein sequences of 16681-PDK53 were also compared with the published sequences of other flaviviruses to obtain a detailed classification of 17 flaviviruses using the neighbor-joining tree method. The analyses of the sequence data produced dendrograms which supported the traditional groupings based on serological evidence, and they suggested that the flaviviruses have evolved by divergent mutational change and there was no evidence of genetic recombination between members of the group. Comparisons of the sequences of the flavivirus polymerase and helicase-like proteins (NS5 and NS3, respectively) with those from other viruses yielded a classification of the flaviviruses indicating that the primary division of the flaviviruses was between those transmitted by mosquitoes and those transmitted by ticks.	
s11756	the use of feline	the use of feline herpesvirus and	1992

id	st_title	st_abstract	st_year
	herpesvirus and baculovirus as vaccine vectors for the gag and env genes of feline leukaemia virus.	baculovirus as vaccine vectors for the gag and env genes of feline leukaemia virus.. The env and gag genes from feline leukaemia virus were expressed in a thymidine kinase-negative feline herpes-virus and a baculovirus. Cats were vaccinated with various combinations of these recombinant viruses and 100% protection against feline leukaemia virus challenge was achieved using an immunization schedule which utilized both env and gag products delivered at both a mucosal and systemic site.	
s11867	relocation of antigens to the cell surface membrane can enhance immune stimulation and protection.	relocation of antigens to the cell surface membrane can enhance immune stimulation and protection.. The major outer capsid glycoprotein of rotaviruses, VP7, is normally synthesized and directed to the ER, where it is required for virus assembly. By substituting a foreign signal sequence for the VP7 signal peptide, a secreted form of VP7 with an authentic amino terminus was produced. Secreted VP7 was further modified by the addition of a transmembrane anchor and cytoplasmic domain to its C-terminus. When the novel chimeric protein was expressed in transfected cells it became anchored in the cell surface membrane. The antigenicity of the chimeric protein was compared with that of the intracellular form of VP7 using recombinant vaccinia viruses to deliver the antigens in vivo. The novel antigen produced enhanced stimulation of both B and T lymphocytes of the immune system, and in mice it was able to induce protection against rotavirus-induced	1992

id	st_title	st_abstract	st_year
		diarrhoeal disease. Other secreted and intracellular antigens show a similar improved level of antigenicity as a result of their relocation to the cell surface. Surface localization may therefore have general utility in the development of recombinant subunit vaccines.	
s12450	blocking elisa for distinguishing infectious bovine rhinotracheitis virus (ibrv)-infected animals from those vaccinated with a gene-deleted marker vaccine.	blocking elisa for distinguishing infectious bovine rhinotracheitis virus (ibrv)-infected animals from those vaccinated with a gene-deleted marker vaccine.. A sensitive and specific blocking enzyme-linked immunosorbent assay (ELISA) was developed to distinguish infectious bovine rhinotracheitis virus (IBRV)-infected animals from those immunized with a glycoprotein gIII deletion mutant, IBRV(NG)dltkdlgIII. For this ELISA, undiluted test sera are used to block the binding of an anti-IBRV gIII monoclonal antibody (mAbgIII)-horseradish peroxidase (HRPO) conjugate to gIII antigen. TMB substrate is used for color development. Negative S/N values (defined as the absorbance at 650 nm of test sera/absorbance at 650 nm of negative control sera) of > 0.80 were obtained with immune sera from gnotobiotic cattle immunized with several bovine viruses, with bovine antisera to bovine herpesvirus-2, and vesicular stomatitis virus, with porcine antisera to pseudorabies virus and parvovirus, and with normal sera from heterologous species. Negative S/N values were also obtained with sera from rabbits twice vaccinated with IBRV(NG)dltkdlgIII. However, the S/N values became positive (S/N < 0.8) 10 to 17 days after the rabbits	1992

id	st_title	st_abstract	st_year
		were challenge exposed to virulent IBRV(Cooper). Most of 116 sera (84%) from feedlot cattle with virus neutralization (VN) titers of < 1:2 or < 1:4 had negative S/N values > 0.8, but 18 sera with negative VN titers had positive S/N values, consistent with observations indicating that an IBRV outbreak was occurring in one of the feedlot herds. Thirty nine sera (98%) from feedlot cattle with VN titers of 1:2 to 1:128 had positive S/N values (< 0.8). One serum with a VN titer of 1:2 had a borderline (+/-) S/N value of 0.81. After immunization with a commercial gIII-positive IBRV vaccine, 115/116 sera with VN titers of 1:2 to 1:256 had positive S/N values (< 0.8). One serum with a VN titer of 1:2 had a negative S/N value of 0.83. Serum from one vaccinated animal that failed to seroconvert after vaccination (VN < 1:4) showed a strongly positive ELISA S/N of 0.48.	
s12471	multiple amino acids in the capsid structure of canine parvovirus coordinately determine the canine host range and specific antigenic and hemagglutination properties.	multiple amino acids in the capsid structure of canine parvovirus coordinately determine the canine host range and specific antigenic and hemagglutination properties.. Canine parvovirus (CPV) and feline panleukopenia virus (FPV) are over 98% similar in DNA sequence but have specific host range, antigenic, and hemagglutination (HA) properties which were located within the capsid protein gene. In vitro mutagenesis and recombination were used to prepare 16 different recombinant genomic clones, and viruses derived from those clones were analyzed for their in vitro host range, antigenic, and HA properties. The region of CPV	1992

id	st_title	st_abstract	st_year
		from 59 to 91 map units determined the ability to replicate in canine cells. A complex series of interactions was observed among the individual sequence differences between 59 and 73 map units. The canine host range required that VP2 amino acids (aa) 93 and 323 both be the CPV sequence, and those two CPV sequences introduced alone into FPV greatly increased viral replication in canine cells. Changing any one of aa 93, 103, or 323 of CPV to the FPV sequence either greatly decreased replication in canine cells or resulted in an inviable plasmid. The Asn-Lys difference of aa 93 alone was responsible for the CPV-specific epitope recognized by monoclonal antibodies. An FPV-specific epitope was affected by aa 323. Amino acids 323 and 375 together determined the pH dependence of HA. Amino acids involved in the various specific properties were all around the threefold spikes of the viral particle.	
s12601	identification and functional analysis of the fowlpox virus homolog of the vaccinia virus p37k major envelope antigen gene.	identification and functional analysis of the fowlpox virus homolog of the vaccinia virus p37k major envelope antigen gene.. A fowlpox virus (FPV) gene with homology to the vaccinia virus p37K major envelope antigen gene was identified and sequenced. The predicted product has a molecular weight of 43,018 Da (p43K). The FPV p43K gene has 37.5% identity with its vaccinia counterpart and higher homology with a molluscum contagiosum virus gene (42.6% identity). Based on upstream sequences, p43K appears to be regulated as a late gene.	1992

id	st_title	st_abstract	st_year
		Recombinant FPV were generated in which a large portion of p43K was replaced by the Escherichia coli lacZ gene. These recombinants failed to produce visible plaques under standard conditions. After prolonged incubation the microplaques developed into small macroscopic plaques. Plaques were purified on the basis of lacZ expression. Single-cycle growth curves comparing the p43K-deleted recombinant (designated fJd43Z) with parental FPV showed that the two viruses produce identical amounts of intracellular virions, but that fJd43Z released 20-fold fewer infectious particles into the medium. CsCl gradient centrifugation of [3H]thymidine-labeled virus was employed to examine differences in the production of physical particles. The two viruses produced equivalent levels of intracellular virions, but fJd43Z failed to produce detectable levels of released particles. FPV p43K is therefore involved in the release of virions from infected cells.	

```
dbRemoveTable(mysqlconnect,'paper')
```

```
## [1] TRUE
```

```
dbRemoveTable(mysqlconnect,'paper_viruse')
```

```
## [1] TRUE
```

```
dbDisconnect(mysqlconnect)
```

```
## [1] TRUE
```