生物信息学学习笔记

张洋[1]

最后更新日期：Friday 16th June, 2017，10:21

[1]邮箱：hiseq@outlook.com

目录

第一部分

常用数据库说明

第一章　cBioPortal 数据库简介

cBioPortal 数据库主要针对癌症基因组学（Cancer Genomics）研究，其提供了大规模癌症基因组的可视化、分析和下载功能。其官方网址为：http://www.cbioportal.org/

1.1　API 简介

通过 cBioPortal 提供的 CGDS 网络服务（Cancer Genomic Data Server web service），可以通过编程的方式快速获取 cBioPortal 所有的基因组学数据。CGDS 网络服务是基于 REST 的，其返回的数据是 Tab 键分隔的文本格式或者 XML 格式。可以选择的编程语言包括：

- Python
- R
- Perl
- Java
- MatLab

所有的请求通过http://www.cbioportal.org/webservice.do提交，请求时需要附上一些必要的参数：

- cmd：需要执行的操作，可选值包括：
  - getTypesOfCancer：获取癌症类型。
  - getNetwork：获取。
  - getCancerStudies：获取癌症研究。
  - getGeneticProfiles：获取特定癌症研究项目的 Genetic Profile 信息。

  - getProfileData：获取一个或多个基因的 genomic profile 数据。
  - getCaseLists：获取。
  - getClinicalData：获取样本的临床信息。
  - getMutationData：获取基因的突变信息。

- 其他的一些可选参数，该参数跟所执行的 cmd 相关。

比如,我们可以通过以下链接获取 cBioPortal 中所有的有关癌症的研究项目： http://www.cbiopor

对于不同的 cmd，其所对应的可选参数也不同。

1.1.1 getTypesOfCancer 命令

getTypesOfCancer 命令用于获取服务器中存储的癌症列表。在调用该命令时，不需要可选参数。其返回的数据包括两列：

type_of_cancer_id	cBioPortal 中唯一表示该癌症的编号。
name	癌症名称

1.1.2 getCancerStudies 命令

getCancerStudies 命令可以用来获取服务器上存储的有关癌症的研究项目的基础数据。在调用该命令时，不需要可选参数。其返回的数据包括三列：

cancer_study_id	cBioPortal 中唯一表示该癌症研究项目的编号。
name	研究项目的名称
description	有关该研究项目的简单说明

1.1.3 getGeneticProfiles 命令

getGeneticProfiles 命令用于获取某个癌症研究项目的所有元数据，包括变异信息、拷贝数信息等。在调用该命令时需要提供一个可选参数：

cancer_study_id	癌症研究项目的编号。

其返回的数据包括六列：

genetic_profile_id	cBioPortal 中唯一表示该 genetic profile 的编号。
genetic_profile_name	genetic profile 的名称
genetic_profile_description	genetic profile 的简单说明
cancer_study_id	癌症研究项目的 ID
genetic_alteration_type	有关该研究项目的简单说明
show_profile_in_analysis_tab	有关该研究项目的简单说明

举例

通过链接：

http://www.cbioportal.org/webservice.do?cmd=getGeneticProfiles&cancer_study_id=msk_impact_2017

可以获取癌症研究项目“MSK-IMPACT”的 Genetic Profiles。

1.1.4 getCaseLists 命令

getCaseLists 命令可以返回特定癌症研究项目中的样本信息。在调用该命令时需要提供一个必选参数：

cancer_study_id	癌症研究项目的编号。

其返回信息包括五列：

case_list_id	
case_list_name	
case_list_description	
cancer_study_id	
case_ids	

1.1.5 getProfileData 命令

getProfileData 命令可以返回一个或多个基因的 genomic profile 数据。在调用该命令时需要提供三个必选参数：

case_set_id	由getCaseLists命令返回的 Case Set 的 ID。
genetic_profile_id	由getGeneticProfiles命令返回的 Genetic Profile。
gene_list	基因列表，多个基因之间以英文逗号“,”分隔。

1.1.6 getMutationData 命令

getMutationData 命令可以获取一些额外的信息，比如变异的注释信息。使用该命令时，需要添加以下参数：

genetic_profile_id	必选	由getGeneticProfiles命令返回的 *genetic_profile_id*。
case_set_id	可选	由getGeneticProfiles命令返回的 *case_list_id*。
gene_list	必选	以逗号分隔的基因列表（HUGO 基因名称或 Entrez 的基因 ID）。

1.1.7 getClinicalData 命令

getClinicalData 用于获取癌症研究项目中所涉及的样本的基本临床信息。可以获取到的基本信息包括样本编号（CASE_ID）癌症类型（CANCER_TYPE）、癌症稍微详细的描述（CANCER_TYPE_DETAILED）、所使用的 DNA 量（DNA_INPUT）、癌症的转移位置（METASTATIC_SITE）、Oncotree 编号（ONCOTREE_CODE）、原发灶（PRIMARY_SITE）、样本类型（SAMPLE_CLASS）、患者性别（SEX）等。在调用该命令时需要提供一个可选参数：

case_set_id	样本集编号

1.2 cBioPortal 使用流程

cBioPortal 数据下载流程：

1. 通过“getCancerStudies”命令，获取癌症研究项目的基本信息，主要获取 *cancer_study_id* 值；
2. 以 *cancer_study_id* 值为参数，调用“getCaseLists”命令，获取癌症项目研究的样本集合的 *case_list_id* 值。
3. 以 *case_list_id* 值作为参数，调用“getClinicalData”命令，获取癌症项目研究样本的临床信息。
4. 以 *cancer_study_id* 值为参数，调用“getGeneticProfiles”命令，获取癌症项目研究的 Genetic Profile 的列表。
5. 以 *genetic_profile_id* 为参数，*case_list_id* 值作为 *case_set_id* 的参数

第二部分

Bioinformatics-related File Formats

第二章　SAM Format

2.1 Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

SAMtools is hosted by GitHub. The project page is here. The source code releases are available from the download page. You can check out the most recent source code with:

```
git clone git://github.com/samtools/samtools.git
```

第三章　生物信息学常用软件

3.1 MutationAssessor：预测氨基酸变异对蛋白功能的影响

该工具可以通过网址：`http://mutationassessor.org/`来访问。

This server predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The functional impact is assessed based on evolutionary conservation of the affected amino acid in protein homologs. The method has been validated on a large set (60k) of disease associated (OMIM) and polymorphic variants. To explore the functional impact of missense mutations found in The Cancer Genome Atlas please use cBioPortal for Cancer Genomics.

3.1.1 How it Works

This server provides semantic linking to variant analysis, annotations, variant multiple sequence alignment html page, and variant 3D structure page.

Please note that the analysis of submitted variations is done asynchronously - if a new variant falls into a protein domain which does not yet have a multiple sequence alignment (MSA) in the server database, "word [sent]" is returned in the "MSA" field until the MSA is built. You can see the size of current MSA queue on about page. The same approach applies when computing Functional Impact scores of new variants.

Input

The server accepts list of variants, one variant per line, plus optional text describing your variants, in genomic coordinates, "+" strand assumed :

`<genome build>,<chromosome>,<position>,<reference allele>,<substituted allele>`

Genome build is optional (build 19 assumed), accepted values: 'hg19' and 'hg38'

Examples:

```
hg38,13,32338418,G,T   BRCA2
hg19,7,55211080,G,A   EGFR
7,55211080,G,A   EGFR
```

or in protein space:

```
<protein ID> <variant> <text>,
```

where <protein ID> can be :

- - Uniprot protein accession (e.g. EGFR_{HUMAN})$-NCBIRefseqproteinID(e.g.NP_0$05219)

Examples:

```
EGFR_HUMAN R521K
EGFR_HUMAN R98Q Polymorphism
EGFR_HUMAN G719D disease
NP_000537 G356A
NP_000537 G360A dbSNP:rs35993958
NP_000537 S46A Abolishes phosphorylation
```

ID types can be mixed in one list in any way.

The server maps each variant to both Uniprot and Refseq protein sequences (if possible).

If the reference residue in the Uniprot protein sequence is different from the one indicated in your variant the analysis will not be performed.

For non-human variants please use Uniprot IDs as mapping to Refseq is not supported.

Uniprot IDs are used to extract information about domain boundaries (Pfam, Uniprot), annotated functional regions (Uniprot), protein-protein interactions (Piana). Refseq protein IDs are used to extract known alterations in cancer (COSMIC), SNPs (dbSNP) and known role in cancer (CancerGenes).

The server determines domain boundaries (using Pfam or Uniprot) for the region with the variant and builds multiple sequence alignment using all Uniprot protein sequences or uses existing one from the repository. To obtain the list of existing alignments in the repository for a giver protein please see WEBAPI section below.