

Getfasta 软件使用说明书

1. 引言

Getfasta 软件基于美国国家生物信息中心(NCBI)开发的 Entrez Direct (E-utilities on the UNIX Command Line) 工具(Kans, 2018), 批量下载 GenBank 数据库中的基因序列数据, 并将结果保存为用于生物信息学分析的 fasta 格式文件。

1.1 编写目的

本说明书为在 Linux/UNIX 环境下使用 Getfasta 软件的用户编写。它将指引使用者按步骤搭建该软件的运行环境, 明确输入文件(Genbank 收录号列表)的录入格式, 熟悉运行参数的配置规则, 理解软件运行时出现的状态信息并掌握获取输出 fasta 格式文件的方法。

1.2 项目背景

GenBank 是美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)建立的 DNA 序列数据库, 为全世界科研人员提供测序获得的基因序列。该数据可通过 NCBI 开发的数据库检索系统(Entrez)对 GenBank 进行访问。

在 Entrez 中, Entrez Direct (E-direct) 是基于 Linux/UNIX 命令行界面检索、下载数据的组件。但该组件仅支持基于单个 GenBank 收录号检索序列信息, 不支持通过收录号列表批量获取序列数据。

本项目开发的 Getfasta 软件能基于 GenBank 收录号列表文件自动化批量下载 GenBank 序列数据, 并自动生成用于生物信息学分析的 fasta 格式文件。

1.3 定义(专门术语的定义和缩写词的原意)

- GenBank 收录号或索引号 (GenBank Accession Numbers): 科研人员将测序获得的核酸或蛋白质序列提交至 GenBank 数据库后, 由 NCBI 接收并分配给该序列的登记号。
- aclist 格式 (aclist format): aclist 是一种使用 UTF-8 编码的文本文件, 是 Getfasta 软件的输入文件格式。该文件存储了需要批量下载序列信息的 GenBank 收录号列表, 列表间用半角逗号 “,” 分隔。
- ft2 格式 (ft2 format): ft2 是一种使用 UTF-8 编码的文本文件, 是为了获得 fasta 格式文件而生成的中间文件。
- fasta 格式 (fasta format): fasta 格式是一种基于文本用于表示核酸序列或多肽序列的格式。其中核酸或氨基酸均以单个字母来表示, 且允许在序列前添加序列名及注释。该格式已成为生物信息学领域常用的

标准文件格式。

- *xx*.aclist 收录号列表文件
- *n* 待下载收录号总数
- *n_d* 已下载的序列总数
- 5 ● *n_{fd}* 全序列均为指定基因区域的序列数
- *n_{pd}* 部分序列为指定基因区域的序列数
- *n_i* 不含指定基因区域的序列数
- *n_f* 下载失败的序列数
- **X** gene 指定基因区域名称
- 10 ● *Locus*_{1-n} 收录号
- *Taxon*_{1-n} 物种名
- *x*_{1-n} 所含基因
- *l*_{1-n} 指定基因区域序列长度
- *L*_{1-n} 全序列长度
- 15 ● *s*_{1-n} 指定基因区域在全序列中的起始位点
- *e*_{1-n} 指定基因区域在全序列中的终止位点
- *t*_{1-n} 下载时间

20 2. 软件概述

2.1 目标

自动化批量获取多物种的全基因组（或同源基因）序列，用于核酸序列一致性比对或分子系统发育关系推断等生物信息学分析。

2.2 功能

- 25 根据输入的收录号列表文件批量下载 GenBank 数据库中的全基因组（或同源基因）的核酸序列，并自动生成 fasta 格式文件。

2.3 性能

a. 数据精确度

- 30 本软件不涉及数字的计算和处理，输入、输出数据的格式均为文本类型。

b. 时间特性

- 35 本软件下载单条核酸序列的所需时间由序列长度、运行软件主机的性能、网络速度和 NCBI 服务器响应时间的决定。本软件对输入数据、输出数据的处理时间由收录号列表长度、核酸序列长度和运行软件主机性能决定。在运行本说明书的示例时，如使用阿里云云服务器实例

(实例规格: ecs.t5-1c1m1.small<北京区域>), 输入文件读取时间<1s, 序列的下载时间<20s/条, fasta 格式文件生成时间<20s。

c.灵活性【在操作方式、运行环境需做某些变更时软件的适应能力。】

5 3. 运行环境

3.1 硬件

推荐配置:

CPU: 主频 > 2.5 GHz (x86 架构)

内存: > 1 GB (DDR4)

10 硬盘: 因数据量确定

网络带宽: > 5Mbps (峰值)

3.2 支持软件

操作系统: **Linux (CentOS 7/8) / UNIX**

Windows (使用 Cygwin 模拟器)

15 支持软件: **Perl¹** (版本号>= v5.30.1)

(须安装以下模块:

warnings; File::Spec; lib; IO::Socket::SSL; Data::Dumper;

Encode; Getopt::Long; HTML::Entities; JSON::PP;

LWP::Simple; LWP::UserAgent; LWP::Protocol::https;

20 MIME::Base64; Net::FTP; Net::hostent; POSIX;

Time::HiRes; URI::Escape; XML::Simple)

Entrez Direct² (版本号>= v13.5)

4. 使用说明

25 4.1 安装和初始化

(1) Linux 系统 (CentOS 7/8)

执行以下脚本自动化完成 Getfasta 和支持软件的安装:

wget

https://ifigure.oss-cn-beijing.aliyuncs.com/SHELL/Getfasta/setup_shell/Getfasta_setup.CentOS.sh

30

chmod +x Getfasta_setup.CentOS.sh

. Getfasta_setup.CentOS.sh

(2) Windows (使用 Cygwin 模拟器)

¹ Perl: <https://www.perl.org/>

² Entrez Direct: E-utilities on the UNIX Command Line: <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

a) 安装 Cygwin 模拟器（附件 1）

b) 执行以下脚本自动化完成 Getfasta 和支持软件的安装：

```
wget
```

```
https://ifigure.oss-cn-beijing.aliyuncs.com/SHELL/Getfasta/setup\_shell/Getfasta\_setup.Cygwin.sh
```

```
chmod +x Getfasta_setup.Cygwin.sh
```

```
. Getfasta_setup.Cygwin.sh
```

4.2 输入文件

Getfasta 的输入数据为下载序列的 GenBank 收录号列表。此列表须使用 aclist 格式文件存储。

4.3 输出文件

Getfasta 的输出数据包括*.ft2、*.fasta 和*.log(可选)共 3 个文件。

ft2 文件：ft2 文件是为了自动化生成 fasta 格式文件而生成的中间文件。该文件存储了根据 GenBank 收录号检索到的该序列详细信息，包括序列号、序列全长、物种名、物种收录号（Taxon ID）、指定基因区域的起始位点、指定基因区域的终止位点、指定基因名称等。

fasta 文件：根据输入的 GenBank 收录号列表自动生成的 fasta 格式文件。

log 文件。当使用 tee 管道命令时，log 文件存储了 Getfasta 运行时所有的屏幕提示信息。

4.4 出错和恢复

出错信息 1：“The accession number is invalid or download failure.”。

含意：因用户提供的收录号有误或网络错误，Getfasta 无法下载数据。

措施：（1）检查收录号是否正确；

（2）将所有下载失败的收录号重组成新的 aclist 文件，再次运行 Getfasta 直至下载成功。若多次下载失败，请用户检查网络是否畅通。

出错信息 2：“提示符光标闪烁，程序无响应”。

含意：用户未输入-f 参数或提供的 aclist 文件有误。

措施：按 Ctrl+C 组合键强制退出程序，重新输入正确的-f 参数或加载合法的 aclist 格式文件。

出错信息 3：“提示符光标不动，程序无响应”。

含意：程序运行异常。

措施：按 Ctrl+C 组合键强制退出程序或强制重启 Shell 管理器。

4.5 求助查询

在使用中有任何疑问或发现程序中的 bug，请通过电子邮件（zhanyj@cnu.edu.cn）与作者联系。

5

5. 运行说明

5.1 运行表

```
Getfasta <-f xx.aclist > [-g "X gene" ] [-t TAXON/TID] [| tee xx.log]
```

参数释义：

- 10 -f xx.aclist : 提交收录号列表文件名及路径（必填项）。
-g " X gene ": 指定下载基因区域序列的名称（可选项）。
-t TAXON/TID: 指定生成 fasta 文件的每条序列前的标签类型（可选项）。
当使用“-t”参数时，可以每条下载序列的物种名（TAXON）或物种登记号（Taxon ID, TID）为 fasta 文件中的序列标签。如命令行中无“-t”
15 参数，Getfasta 默认使用每条下载序列的收录号作为 fasta 文件中的序列标签。

| tee xx.log: 保存日志文件的管道指令。

tee 为 Linux/Unix 的内部命令，通过管道操作符 “|” 获取软件运行时的屏幕提示信息并保存在以 log 为扩展名的日志文件中。

20

5.2 屏幕提示信息解析

（0）初始化

显示启动信息，确认运行参数。

25 GetFasta V1.0 ①

The program is designed for retrieving DNA sequences data from GenBank

② automatically.

Author: Yongjia Zhan ③

E-mail: zhanyj@cnu.edu.cn

30 ④

Parameters:

-f [path]/[filename].aclist --input the accession list file(Required). ⑤

-g [gene region] -- input the name of specific gene region in GenBank.

-t Select the label type of fasta file TAXON/TID.

35 The default label type of fasta file is Accession number if the arg '-t'

is
undefined.

Press Ctrl+C to exit the program.

⑥

5 -----

注释:

①【软件名(版本号)】②【功能简介】③【作者姓名】④【E-mail】
⑤【运行参数表】⑥【运行异常处理方法】

10 (1) 读取并解析输入文件

解析输入文件和运行参数信息: 计算待下载序列数量, 确认指定基因区域名称和自定义 fasta 文件序列标签类型。

STEP1: Confirming parameters:

15 Accession number list: **xx**.aclist (Total: **n** accession numbers)

Gene region: **X** gene

The label type of fasta file:

20 (2) 历史结果检查

检查当前目录下是否存在历史结果文件, 如存在则覆盖所有历史结果文件。

STEP2: Checking for previous results...

25 No previous results. / The previous results will be override.

(3) 下载序列信息

(a) 下载全序列。

30 Getfasta -f **xx**.aclist

每条序列下载成功后屏幕提示如下信息:

<序号/序列总数> <收录号>. <物种名称> >>> 此全序列为全基因组
或为包含多个基因的序列(<全序列长度>)/此序列包含<x>基因(<全序列
长度>)/ 收录号错误或下载失败 (<下载时间>).

35 -----

STEP3: Getting sequences based on the list of accession number. Please waiting:

Warning: The current setting did not specified any gene region.

1/ *n Locus₁. Taxon₁* >>> The sequence seems to a complete genome or containing multiple genes (*L₁* bp) / The sequence only contains complete/partial cds of **X** gene (*L₁* bp) / The accession number is invalid or download failure. (Time: *t₁* s).

2/ *n Locus₂. Taxon₂* >>> The sequence seems to a complete genome or containing multiple genes (*L₂* bp) / The sequence only contains complete/partial cds of **X** gene (*L₂* bp) / The accession number is invalid or download failure. (Time: *t₂* s).

.....

n/ n Locus_n. Taxon_n >>> The sequence seems to a complete genome or containing multiple genes (*L_n* bp) / The sequence only contains complete/partial cds of **X** gene (*L_n* bp) / The accession number is invalid or download failure. (Time: *t_n* s).

(b) 下载指定基因区域的序列。

Getfasta -f *xx.aclist* -g "**X** gene"

每条序列下载成功后屏幕提示如下信息:

<序号/序列总数> <收录号>. <物种名称> >>> 此全序列仅包含 **X** 基因(<全序列长度>) /部分序列包含 **X** 基因 <**X** 基因序列长度>(<全序列长度>): <起始位点>, <终止位点> /此序列不包含 **X** 基因, 忽略下载! / 收录号错误或下载失败 (<下载时间>).

STEP3: Getting sequences based on the list of accession number. Please waiting:

1/ *n Locus₁. Taxon₁* >>> The sequence only contains complete/partial cds of **X** gene (*L₁* bp) / The partial sequence contains **X** gene. **X** gene length: *l₁* bp (*L₁* bp), Start: *s₁*, Stop: *e₁*/ The sequence does not contain the sequence of **X** gene. Ignored / The accession number is invalid or download failure. (Time: *t₁* s).

2/ *n Locus₂. Taxon₂* >>> The sequence only contains complete/partial cds of **X** gene (*L₂* bp) / The partial sequence contains **X** gene. **X** gene length: *l₂* bp (*L₂* bp), Start: *s₂*, Stop: *e₂*/ The sequence does not contain the sequence of **X** gene. Ignored! / The accession number is invalid or download failure. (Time: *t₂* s).

.....

n/ n Locus_n. Taxon_n >>> The sequence only contains complete/partial cds of **X**

gene (L_n bp) / The partial sequence contains **X** gene. **X** gene length: l_n bp (L_n bp),
 Start: s_n , Stop: e_n / The sequence does not contain the sequence of **X** gene. Ignored!
 / The accession number is invalid or download failure. (Time: t_n s).

4. 结果汇总

4.1 下载全序列

Getfasta -f xx.aclist

SUMMARY 部分包括序列下载结果汇总和 fasta 文件生成状态提示。

SUMMARY:

Getting n_d sequences in file: xx.aclist.ft2, including: ①

Downloaded failure sequences: n_f ②

Congratulations! ③

The fasta file generated! Getting n_d sequences totally.

注释:

①【已下载的序列总数 n_d 】②【下载失败的序列数 n_f 】③【fasta 文件状态提示】

$$n = n_d + n_f$$

4.2 下载指定基因区域的序列

命令行: Getfasta -f xx.aclist -g "X gene"

SUMMARY 部分包括序列下载结果汇总和 fasta 文件生成状态提示。

SUMMARY:

Getting n_d sequences in file: xx.aclist.ft2, including: ①

the sequences only contain complete/partial cds of **X** gene: n_{fd}

the partial sequences contain **X** gene: n_{pd}

Ignoring other gene sequence: n_i ②

Downloaded failure sequences: n_f ③

Congratulations! ④

The fasta file generated! Getting n_d sequences totally.

注释:

①【已下载的序列总数 n_d 】(含 a.全序列均为指定基因区域的序列数

n_{fd} 和 b.仅部分序列为指定基因区域的序列数 n_{pd}) ②【被忽略的其他不含指定基因区域的序列数 n_i 】③【下载失败的序列数 n_f 】④【fasta 文件状态提示】

$$n = n_d + n_i + n_f ; \quad n_d = n_{fd} + n_{pd};$$

5

9. 用户操作举例

9.1 示例

下载 GenBank 中收录的 2019-新型冠状病毒 (COVID-19 or SARS-COV-2) a).全基因组序列, b) S 基因序列和 c). 2019-新型冠状病毒和其他 47 种 beta 冠状病毒 (*Betacoronavirus*) 全基因组序列(Wu et al., 2020)。

(1) 研究目的:

a) 对 2019-新型冠状病毒的全基因组序列进行核酸一致性比较。

b) 对 2019-新型冠状病毒的 S 基因序列进行分析。该基因编码病毒包膜上的 S 蛋白。该蛋白具有种特异性能与呼吸道黏膜细胞表面的 ACE2 受体蛋白特异性结合, 使病毒进入人体细胞。

c) 2019-新型冠状病毒的分子系统发育分析

(2) 命令行:

a) `Getfasta -f 2019nCoV.aclist | tee 2019nCoV.log`

b) `Getfasta -f COVID-19.aclist -g "S" | tee COVID-19.log`

c) `Getfasta -f betaCoV.aclist -t TAXON | tee betaCoV.log`

(3) 运行提示信息

GetFasta V1.0

The program is designed for retrieving DNA sequences data from GenBank automatically.

Author:Yongjia Zhan

E-mail:zhanyj@cnu.edu.cn

Parameters:

`-f [path]<filename>.aclist --input the accession list file or path(Required).`

`-g [gene region] -- input the name of specific gene region in GenBank.`

`-t Select the alternative label type of fasta file: TAXON/TID.`

The default label type of fasta file is the accession number if the

arg '-t' is undefined.

Press Ctrl+C to exit the program.

a) -----

5 STEP1: Confirming parameters:

Accession number list: 2019nCoV.aclist (Total: 26 accession numbers)(Time:0s)

Gene region: gene

The alternative label type of fasta file:

10 -----

STEP2: Checking for previous results...

No previous results.

STEP3: Getting Sequences based on the list of accession number. Please waiting:

15

Warning: The current setting did not specified any gene region.

1/26 MN908947. Severe acute respiratory syndrome coronavirus 2 >>>
The sequence seems to a complete genome or containing multiple genes(29903 bp)(Time:17s).

20 2/26 MN938384. Severe acute respiratory syndrome coronavirus 2 >>>
The sequence seems to a complete genome or containing multiple genes(29838 bp)(Time:5s).

...

25 26/26 MT020881. Severe acute respiratory syndrome coronavirus 2 >>>
The sequence seems to a complete genome or containing multiple genes(29882 bp)(Time:9s).

SUMMARY:

Getting 26 sequences in file: 2019nCoV.aclist.ft2, including:

30 Downloaded failure sequences: 0

Congratulations!

The fasta file generated (Time: 9s)! Getting 26 sequences totally.

35 **b)** -----

STEP1: Confirming parameters:

Accession number list: COVID-19.aclist (Total: 44 accession numbers)(Time:0s)

Gene region: S gene

5 The alternative label type of fasta file:

STEP2: Checking for previous results...

No previous results.

10 **STEP3: Getting Sequences based on the list of accession number. Please waiting:**

1/44 MN908947. Severe acute respiratory syndrome coronavirus 2 >>>
The partial sequence contains S gene. S gene length:3822bp(29903bp),
Start:21563, Stop:25384(Time:42s).

15 2/44 MN970003. The sequence does not contain the sequence of S gene.
Ignored(Time:11s)!

...

15/44 MN975266. Severe acute respiratory syndrome coronavirus 2 >>>
The sequence only contains complete/partial cds of S gene(107 bp)(Time:4s).

20 ...

44/44 MT020881. Severe acute respiratory syndrome coronavirus 2 >>>
The partial sequence contains S gene. S gene length:3822bp(29882bp),
Start:21563, Stop:25384(Time:15s).

25 **SUMMARY:**

Getting 31 sequences in file: COVID-19.aclist.ft2, including:

the sequences only contain complete/partial cds of S gene : 7

the partial sequences contain S gene: 24

Ignoring other gene sequences: 13

30 Downloaded failure sequences: 0

Congratulations!

The fasta file generated (Time: 15s)! Getting 31 sequences totally.

35 **c)** -----

STEP1: Confirming parameters:

Accession number list: betaCoV.aclist (Total: 48 accession numbers)(Time:0s)

Gene region: gene

5 The alternative label type of fasta file:

STEP2: Checking for previous results...

No previous results.

10 STEP3: Getting Sequences based on the list of accession number. Please waiting:

Warning: The current setting did not specified any gene region.

1/48 AF029248. Murine hepatitis virus >>> The sequence seems to a complete genome or containing multiple genes(31357 bp)(Time:8s).

15 ...

13/48 MN908947. Severe acute respiratory syndrome coronavirus 2 >>> The sequence seems to a complete genome or containing multiple genes(29903 bp)(Time:17s).

20 47/48 KJ473816. BtRs-BetaCoV/YN2013 >>> The sequence seems to a complete genome or containing multiple genes(29142 bp)(Time:11s).

48/48 KY417145. Bat SARS-like coronavirus >>> The sequence seems to a complete genome or containing multiple genes(29710 bp)(Time:7s).

SUMMARY:

25 Getting 47 sequences in file: betaCoV.aclist.ft2, including:

Downloaded failure sequences: 1

Congratulations!

The fasta file generated(Time:7s)! Getting 47 sequences totally.

30 (4) 结果解读

a) 基于收录号列表 (2019nCoV.aclist), 成功下载 2019-新型冠状病毒全基因组序列 26 条, 失败下载 0 条。最终生成 fasta 格式文件 (2019nCoV.aclist.fasta)。

35 b) 基于收录号列表 (COVID-19.aclist), 成功下载 2019-新型冠状病毒 S 基因 31 条 (其中: 7 条序列仅包含全长/部分长度 S 基因, 24 条序列

部分长度包含 S 基因), 不含 S 基因序列 13 条, 失败下载 0 条, 最终生成 fasta 格式文件 (COVID-19.aclist.fasta)。

- 5 c) 基于收录号列表 (betaCoV.aclist), 成功下载 beta 冠状病毒属 47 个物种的全基因组序列 (其中 1 种为 2019-新型冠状病毒), 因网络错误失败下载 1 条, 最终生成 fasta 格式文件 (betaCoV.aclist.fasta) 并以病毒物种名作为 fasta 文件种的序列标签。

参考资料

- 10 Kans, J. (2018). Entrez Direct: E-utilities on the UNIX Command Line. [Updated 2018 Jul 30]. In: Entrez Programming Utilities Help. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK179288/>
- Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., . . . Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*. doi:10.1038/s41586-020-2008-3

附件 1 Cygwin 模拟器安装说明

1. 下载并启动安装程序

http://www.cygwin.com/setup-x86_64.exe

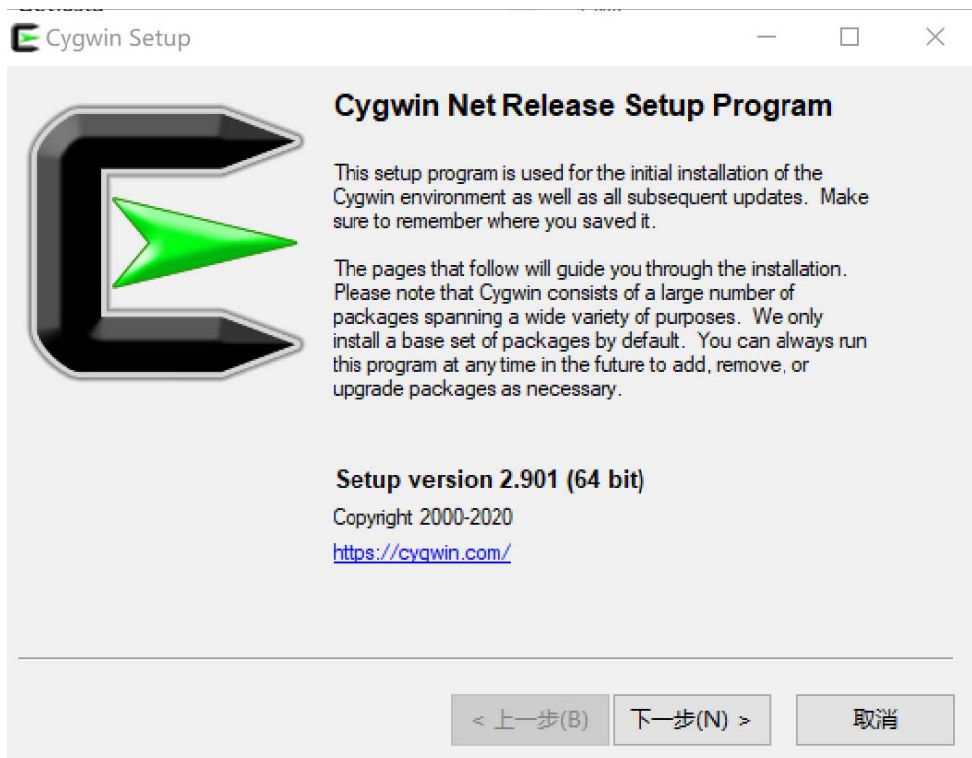
5 or

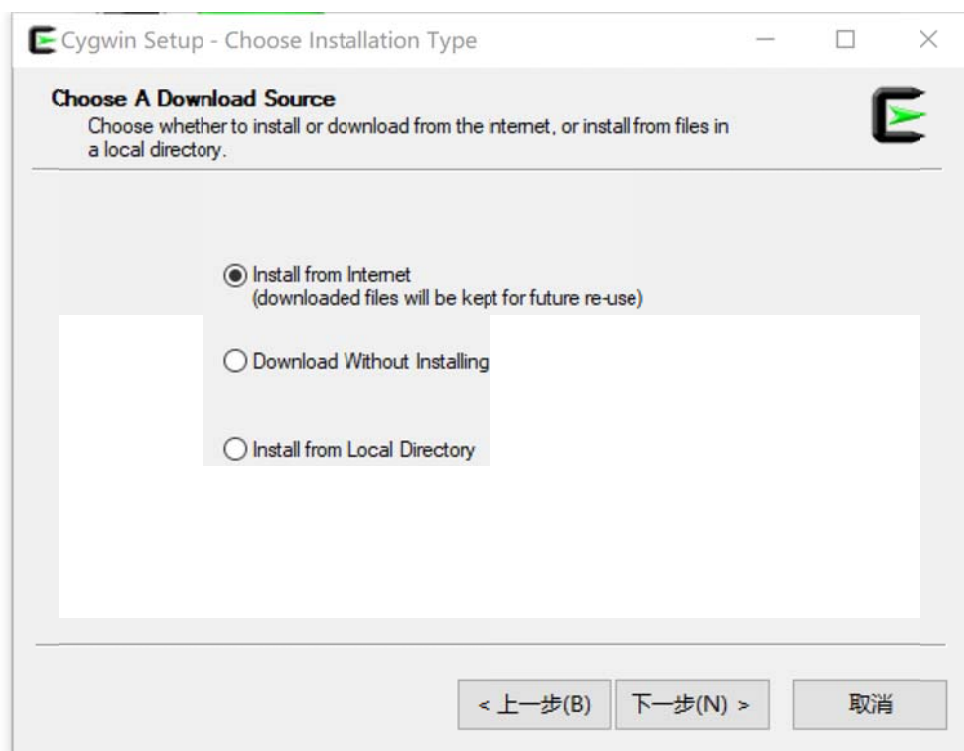
https://ifigure.oss-cn-beijing.aliyuncs.com/SHELL/Getfasta/cygwin_setup-x86_64.exe



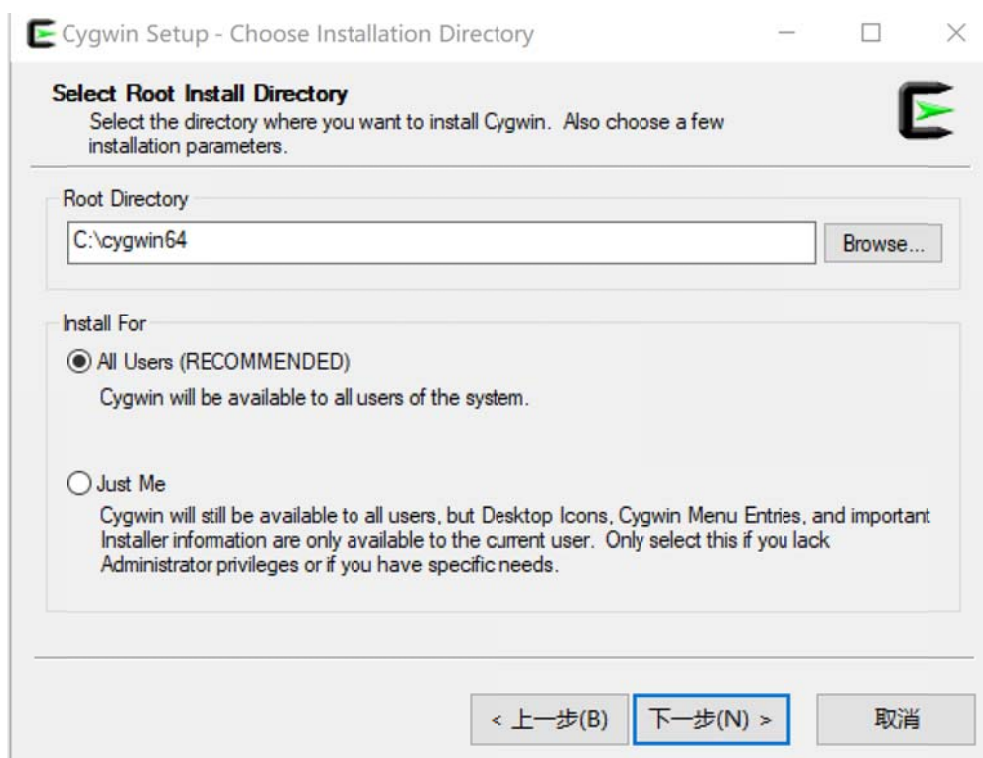
The screenshot shows the Cygwin website. The header features the 'Cygwin' logo and the tagline 'Get that Linux feeling - on Windows'. Below the header, a section titled 'This is the home of the Cygwin project' contains a 'What...' section. This section is divided into two columns: '...is it?' and '...isn't it?'. The '...is it?' column lists that Cygwin is a large collection of GNU and Open Source tools and a DLL (cygwin1.dll) providing substantial POSIX API functionality. The '...isn't it?' column lists that Cygwin is not a way to run native Linux apps on Windows (requiring source rebuild) or a way to magically make native Windows apps aware of UNIX functionality (requiring source build). Below this, a 'Cygwin version' section states the most recent version of the Cygwin DLL is 3.1.4. An 'Installing Cygwin' section instructs users to run 'setup-x86_64.exe' to perform a fresh install or update an existing installation. It also notes that individual packages are updated separately from the DLL.

2. 选择“Install from Internet”模式

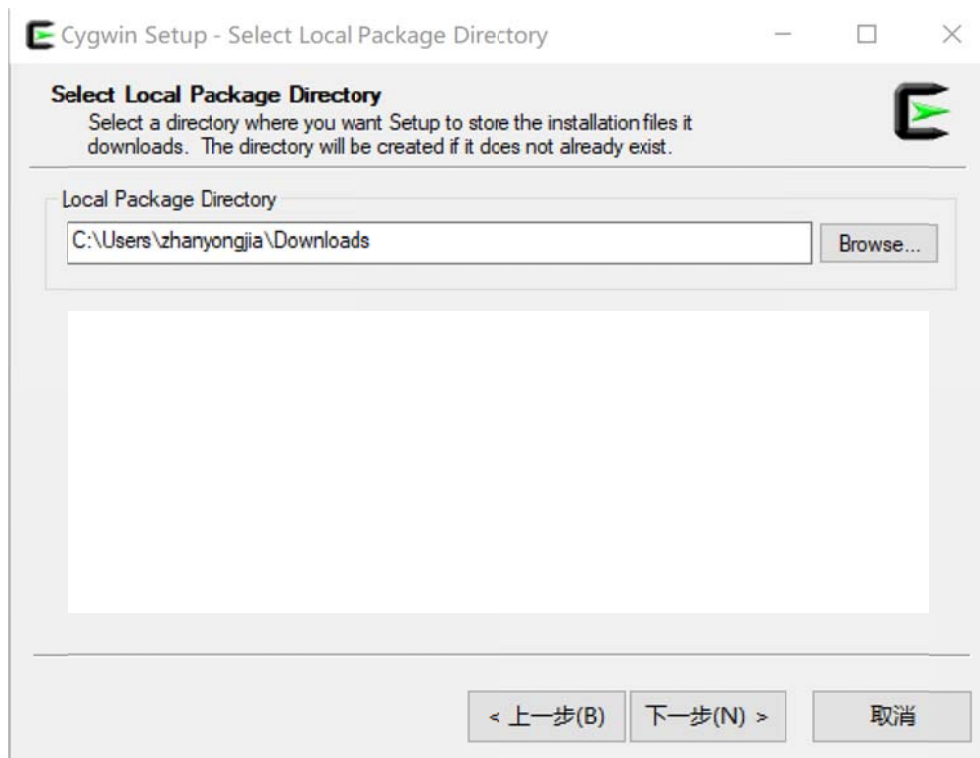




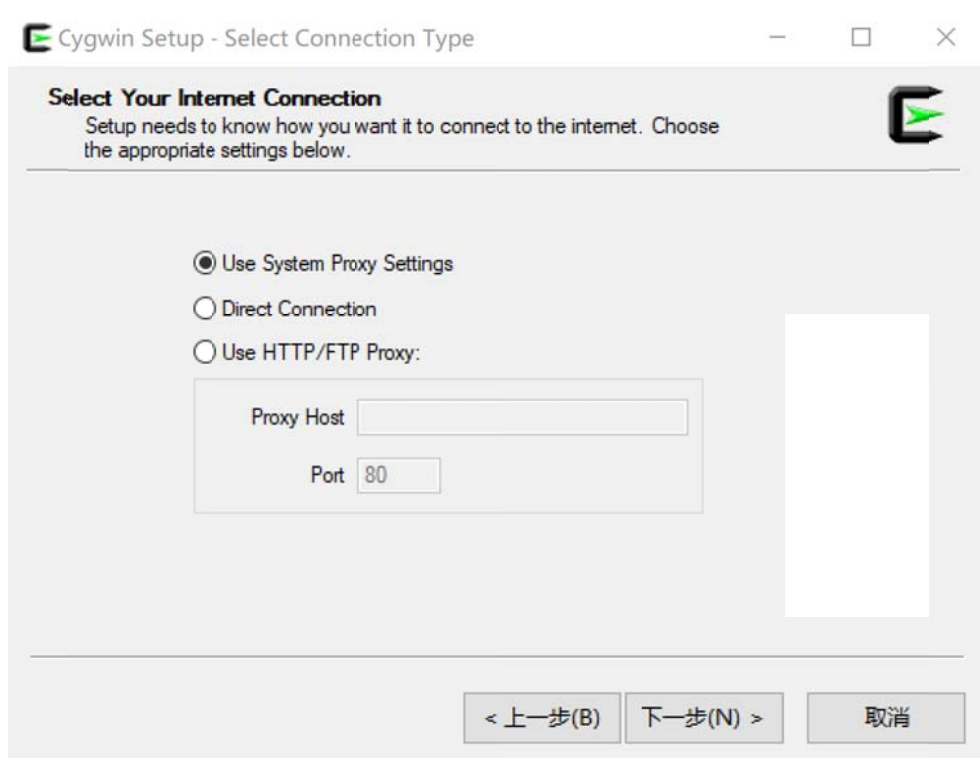
3. 指定程序安装路径



4. 指定程序包安装路径

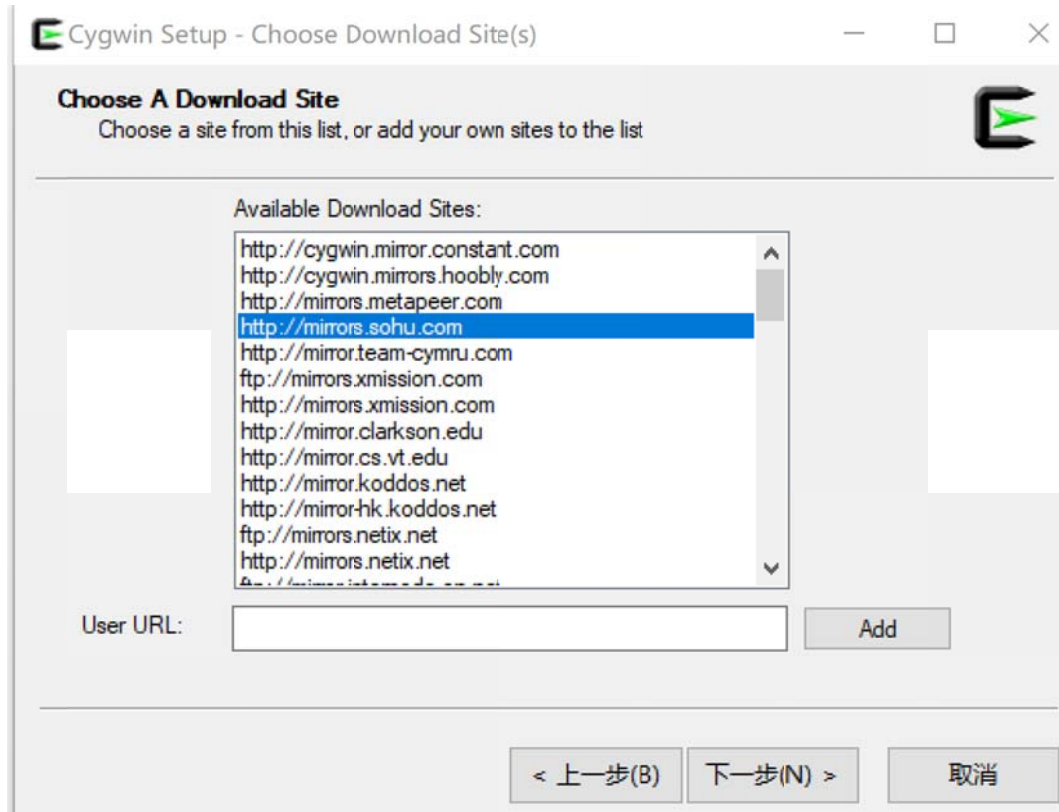


5. 选择 Internet 连接类型



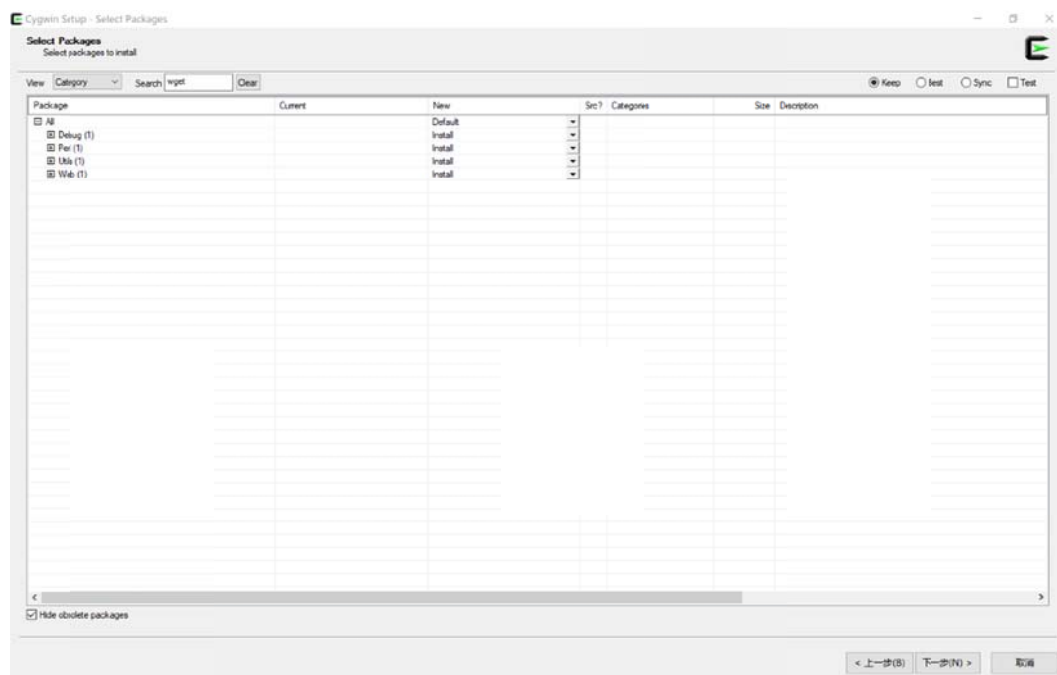
6. 选择安装源镜像

(推荐使用国内源, 如 <http://mirrors.sohu.com/cygwin/>)

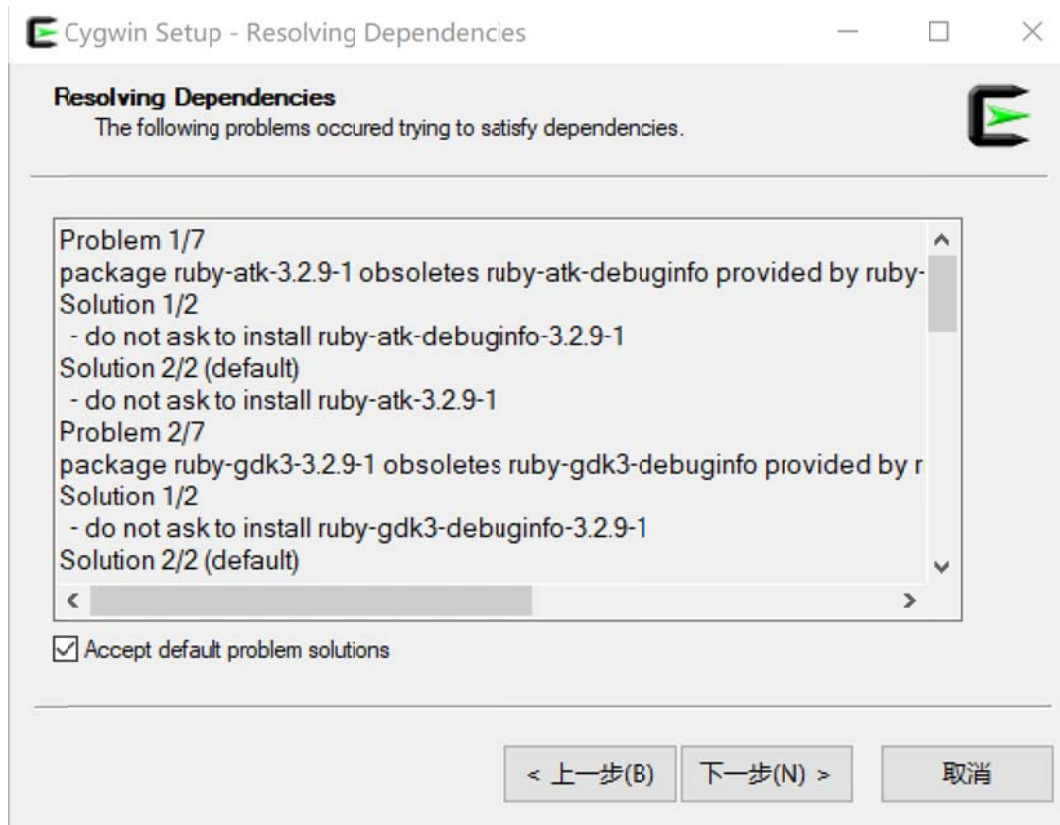


5 7. 安装程序包 wget 和 perl

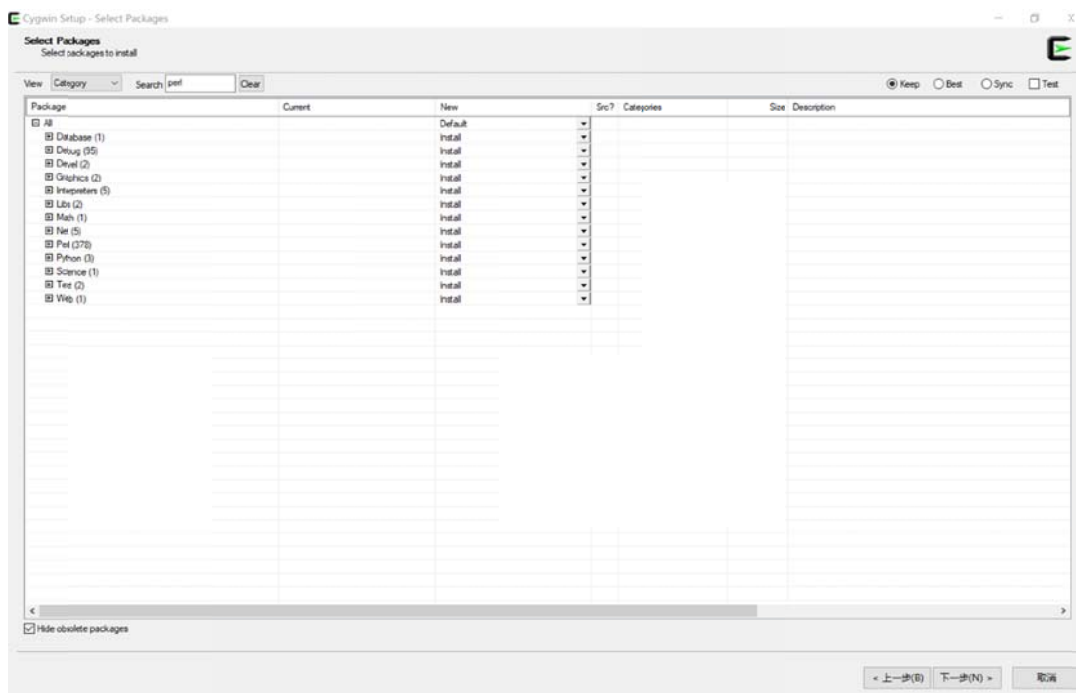
(“View” 处选择 “Category”, “Search” 输入 “wget” / “perl”, 安装模式选择 “keep”, 将 “package” 列表的相关程序包选为 “Install”)

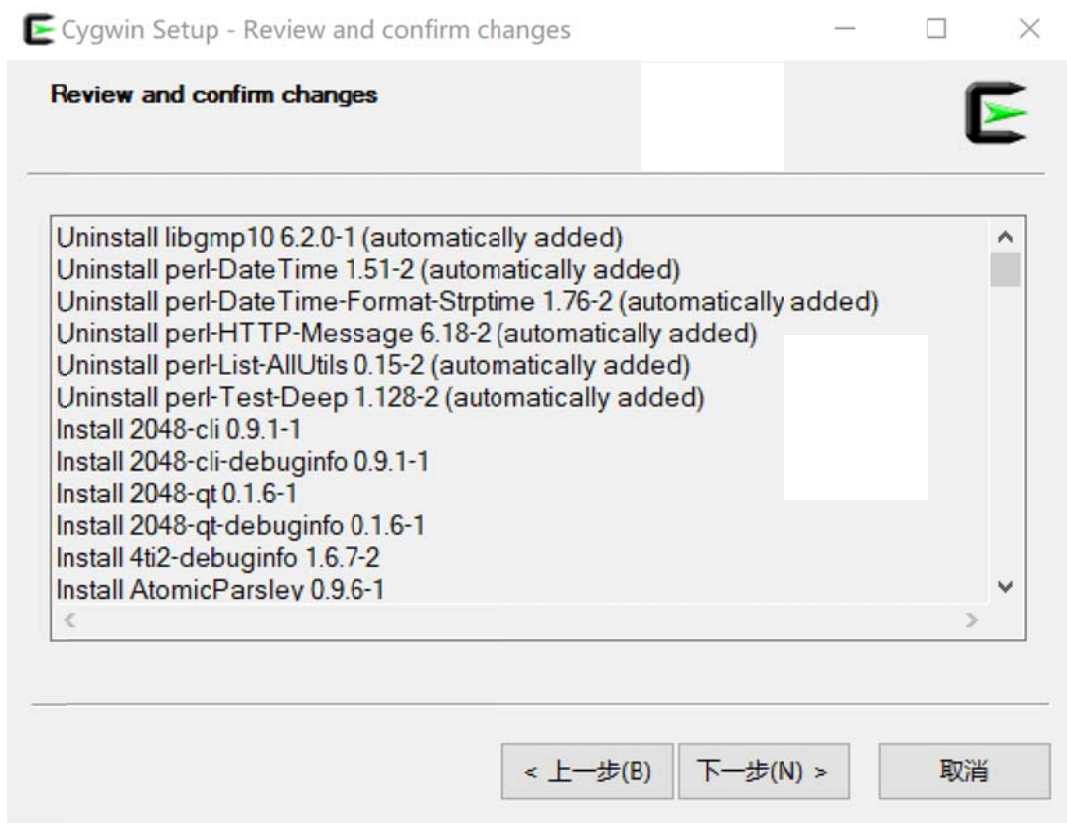


8. 解析程序包依赖关系

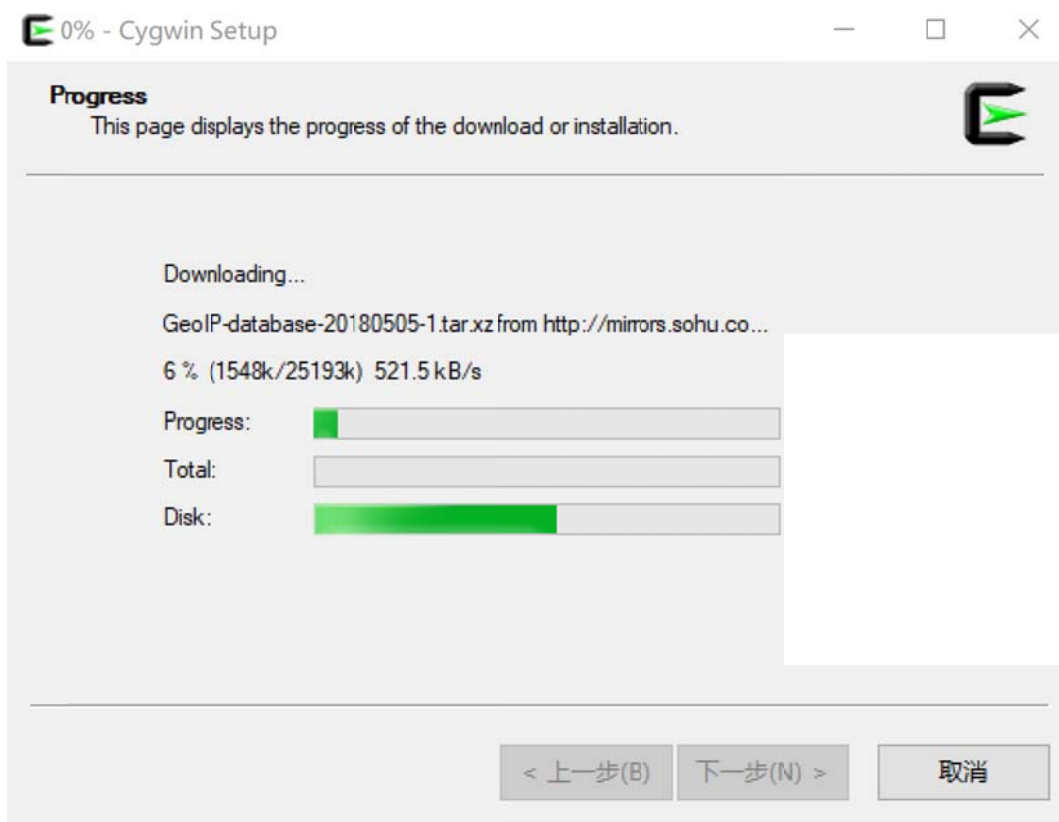


9. 确认安装配置

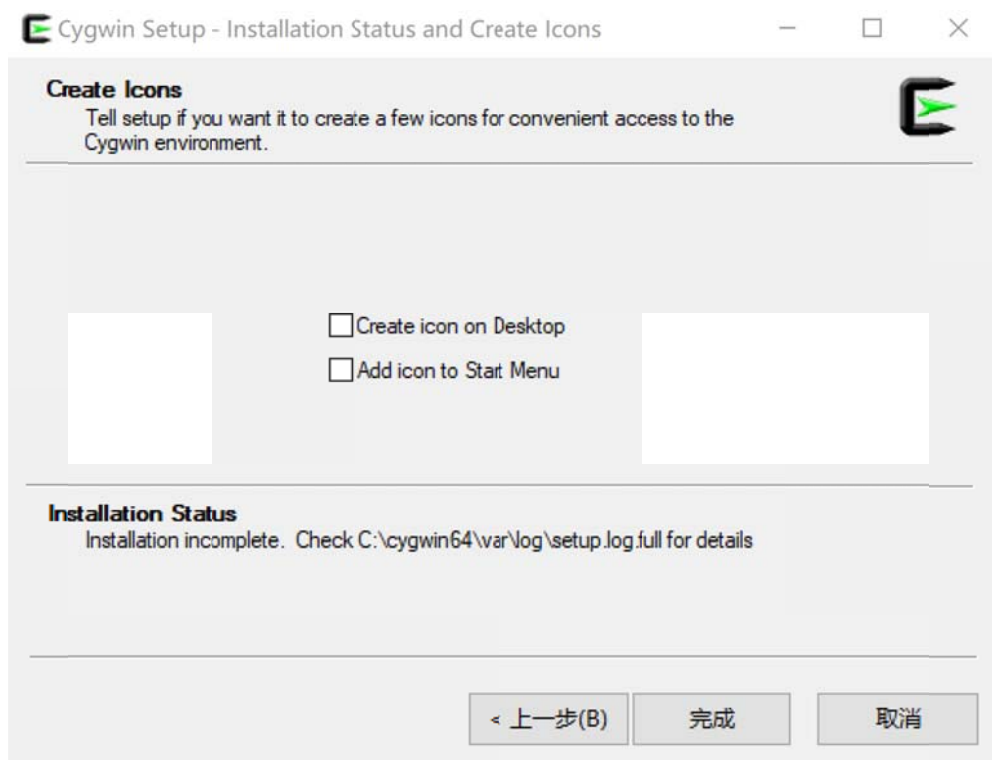




10. 启动安装进程



11. 安装完成



12. 打开 Cygwin，输入“wget”和“perl -v”，确认程序包安装成功

```
zhanyongjia@DESKTOP-HHMUD42 ~  
$ wget  
wget: 未指定 URL  
用法: wget [选项]... [URL]...  
请尝试使用“wget --help”查看更多的选项。  
zhanyongjia@DESKTOP-HHMUD42 ~  
$ perl -v  
This is perl 5, version 30, subversion 1 (v5.30.1) built for x86_64-cygwin-thread-multi  
(with 7 registered patches, see perl -v for more detail)  
Copyright 1987-2019, Larry Wall  
Perl may be copied only under the terms of either the Artistic License or the  
GNU General Public License, which may be found in the Perl 5 source kit.  
Complete documentation for Perl, including FAQ lists, should be found on  
this system using "man perl" or "perldoc perl". If you have access to the  
Internet, point your browser at http://www.perl.org/, the Perl Home Page.  
zhanyongjia@DESKTOP-HHMUD42 ~  
$
```