Gang Chen

- Gang Chen, Ph.D in Computer Science
- VP of BGI-Tech
- Research interests: machine learning and cloud computing for bioinformatics
- chengang@bgitechsolutions.com
- chengangcs@weibo.com
- chengangcs@wechat

C in Bioinformatics

Project and Softwares

Gang Chen
chengang@bgitechsolutions.com

September 20, 2014

华大科技 BGI·tech
香港中文大學 The Chinese University of Hong Kong

Outline

Next

Next

Overview

Smith-Waterman Algorithm

The algorithm is proposed by Temple F. Smith and Michael S. Waterman to perform local sequence alignment.

News

Michael Waterman has become the first honorary professor of the Chinese University of Hong Kong, Shenzhen
http://www.cuhk.edu.cn/News/142.html

Design

I/O

- Input: Two sequences: two string;
- Output: alignment results;

Files

- swa.h: constant and function declarations;
- swa.c: function definitions;
- Makefile: ?

Next

swa.h

see swa/swa.h

swa.c

see swa/swa.c

compile and run

```
gcc swa.c -o swa ./swa
```

Next

GNU Make

GNU Make

GNU Make is a tool which controls the generation of executables and other non-source files of a program from the program's source files.

Website

http://www.gnu.org/software/make/

Makefile for swa

```
1 swa: swa.c swa.h
2         gcc swa.c -o swa -Wall
3 clean:
4         rm swa
```

Tutorials on Makefile

- Tutorial in English: http://makepp.sourceforge.net/
- Official Manual: http://www.gnu.org/software/make/manual/make.html
- 跟我一起写Makefile：http://blog.csdn.net/haoel/article/details/2886
  http://coolshell.cn


ALIBABA GROUP HOLDING LTD (BABA) - NYSE Follow
90.39 22.39 (32.92%) 12:48PM EDT - Nasdaq Real Time Price

Project Management

- Source Version Control: git and github;
- Generation of Makefile: automake
- Documentation: Markdown

Next

Next

1. A project: Sequence Alignment in C
   - Smith-Waterman Algorithm
   - Implementation
   - Project Management: Makefile
2. Bioinformatics Software in C
   - seqtk
   - bwa
   - libsvm

Overview

seqtk

Toolkit for processing sequences in FASTA/Q formats.

Source Codes

https://github.com/lh3/seqtk

Files

- Makefile
- README.md
- khash.h
- kseq.h
- ksort.h
- kstring.h
- ksw.c
- ksq.h
- kvec.h
- seqtk.c
- trimadap.c

Compile and Run

```
make

./seqtk
```

How to read the codes?

Makefile

For most C/C++ projects, the best start point to understand the codes is the Makefile.

Why?

We have to define the relationship between all source files and executable files in the Makefile.

Makefile

```
1  CC=gcc
2  CFLAGS=-g -Wall -O2 -Wno-unused-function
3
4  all:seqtk trimadap
5
6  seqtk:seqtk.c khash.h kseq.h
7                  $(CC) $(CFLAGS) seqtk.c -o $@ -lz -lm
8
9  trimadap:trimadap.c kseq.h ksw.h
10                 $(CC) $(CFLAGS) ksw.c trimadap.c -o $@ -lz -lm
11
12 clean:
13                 rm -fr gmon.out *.o ext/*.o a.out seqtk trimadap *~ *.a *.dSYM session*
```

seqtk.c

see seqtk-master/seqtk.c

Convert FASTQ to FASTA

```
./seqtk seq seq.fq > seq.fa
```

Data

http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR357733

main and stk_seq functions

- main: see seqtk-master/seqtk.c:1371
- stk_seq: see seqtk-master/seqtk.c:1106

Summary

- Makefile can help us understand source codes;
- Typically, declarations are put in header files
- and implementation are put in source files;

Next

1. A project: Sequence Alignment in C
   - Smith-Waterman Algorithm
   - Implementation
   - Project Management: Makefile
2. Bioinformatics Software in C
   - seqtk
   - bwa
   - libsvm

Overview

BWA is a software package for mapping DNA sequences against a large reference genome, such as the human genome. It consists of three algorithms:

- BWA-backtrack:designed for Illumina sequence reads up to 100bp;
- BWA-SW: 70bp to a few megabase;
- BWA-MEM: 70bp to a few megabase, faster and more accurate;

Makefile

see bwa-master/Makefile

main function

- see bwa-master/main.c:60
- Call main_mem: bwa-master/main.c:83
- main_mem declaration: bwa-master/main.c:24

main function

- see bwa-master/main.c:60
- Call main_mem: bwa-master/main.c:83
- main_mem declaration: bwa-master/main.c:24
- Where is the function definition?

main_mem function

- Makefile
- bwamem.c, bwamem.h ...

main_mem function

- Makefile
- bwamem.c, bwamem.h ...
- fastmap.c

Next

1. A project: Sequence Alignment in C
   - Smith-Waterman Algorithm
   - Implementation
   - Project Management: Makefile
2. Bioinformatics Software in C
   - seqtk
   - bwa
   - libsvm

Overview

LIBSVM

LIBSVM is an integrated software for support vector classification(C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM).

Website

http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Overview of Codes

Source Codes

- A mixture of C and C++;
- All C codes are compiled by using C++ compiler;
- C++ is used as a C with class and template;

Makefile

see libsvm-3.18/Makefile

svm-train.c

- main function is definied in svm-train.c
- structs are defined in svm.h
- functions are declared in svm.h and implemented in svm.cpp
- classes and methods are declared and implemented in svm.cpp

svm-train.c

- main function is definied in svm-train.c
- structs are defined in svm.h
- functions are declared in svm.h and implemented in svm.cpp
- classes and methods are declared and implemented in svm.cpp
- the coding style is bad.

C and C++

- C++ is a upgrade of C?
- C++ is C with class?
- C++ is C with class, template, meta-programming and so on?
- What is the relationship between C and C++?

C and C++

- C++ is a new programming languages;
- Currently, C++ is compatible with C;
- The latest version of C is C99, C11 is in progress;
- The latest version of C++ is C++14, C++17 is in progress.

Thanks!