INN700 Literature Review

Genomic Data Searching and Processing

ZUTAO WU

N8975698

Course: IT60 Master by Research

Principal Supervisor: Associate Professor Jim Hogan

Associate Supervisor: Dr Wayne Kelly

Science and Engineering

Queensland University of Technology

September 28, 2015

Abstract

Next Generation Sequencing (NGS) technologies have led to an explosion in genetic and genomic datasets which has posed a significant challenge in existing sequence processing and searching methods, and has motivated new approaches for sequence processing including adaptation of methods from other domains such as information retrieval and deep learning. Measurement of sequence similarity is an important topic in bioinformatics. At present, such process is commonly conducted by BLAST and CLUSTAL, the alignment based similarity determination tools. However, alignment-based approaches may not adapt well on increasing size of datasets as well as on mutation and rearrangement of large gene fragments.

To tackle this problems of genetic data processing and searching, I have reviewed several conventional processing approaches (alignment methods) in bioinformatics, and compare to the alignment free method using locality-sensitive hashing methods. To tackle the problem of pattern and representation learning in bioinformatics, deep learning is a notable topic in such field which may provide us another direction for data processing in bioinformatics.

Contents

List of Figures

List of Tables

1 Test Chinese

测试中文，一二三四五六七

2 Section Headings

We (Ancey et al., 1996) BBB (Radjavi and Rosenthal, 1973) in this section how to obtain headings for the various sections and subsections of our document.

2.1 Headings in the 'article' Document Style

In the 'article' style, the document may be divided up into sections, subsections and sub-subsections, and each can be given a title, printed in a boldface font, simply by issuing the appropriate command.

The foundations[1] of the rigorous study of *analysis* were laid in the nineteenth century, notably by the mathematicians Cauchy and Weierstrass. Central to the study of this subject are the formal definitions of *limits* and *continuity*.

Let D be a subset of $\mathbf{R}$ and let $f\colon D \to \mathbf{R}$ be a real-valued function on D. The function f is said to be *continuous* on D if, for all $\epsilon > 0$ and for all $x \in D$, there exists some $\delta > 0$ (which may depend on x) such that if $y \in D$ satisfies

$$|y - x| < \delta$$

then

$$|f(y) - f(x)| < \epsilon.$$

One may readily verify that if f and g are continuous functions on D then the functions $f + g$, $f - g$ and $f.g$ are continuous. If in addition g is everywhere non-zero then f/g is continuous.

Algorithm 1: IntervalRestriction

Data: $G = (X, U)$ such that G^{tc} is an order.
Result: $G' = (X, V)$ with $V \subseteq U$ such that G'^{tc} is an interval order.
begin
  $V \longleftarrow U$
  $S \longleftarrow \emptyset$
  for $x \in X$ **do**
    $NbSuccInS(x) \longleftarrow 0$
    $NbPredInMin(x) \longleftarrow 0$
    $NbPredNotInMin(x) \longleftarrow |ImPred(x)|$
  for $x \in X$ **do**
    if $NbPredInMin(x) = 0$ **and** $NbPredNotInMin(x) = 0$ **then**
      $AppendToMin(x)$
1 **while** $S \neq \emptyset$ **do**
REM remove x from the list of T of maximal index
2 **while** $|S \cap ImSucc(x)| \neq |S|$ **do**
      for $y \in S - ImSucc(x)$ **do**
        { remove from V all the arcs zy : }
        for $z \in ImPred(y) \cap Min$ **do**
          remove the arc zy from V
          $NbSuccInS(z) \longleftarrow NbSuccInS(z) - 1$
          move z in T to the list preceding its present list
          {i.e. If $z \in T[k]$, move z from $T[k]$ to $T[k-1]$}
        $NbPredInMin(y) \longleftarrow 0$
        $NbPredNotInMin(y) \longleftarrow 0$
        $S \longleftarrow S - \{y\}$
        $AppendToMin(y)$
    $RemoveFromMin(x)$

3 algorithm

4 math

4.1 inline equation

In physics, the mass-energy equivalence is stated by the equation $E = mc^2$, discovered in 1905 by Albert Einstein.

4.2 independent

The mass-energy equivalence is described by the famous equation

$$E = mc^2$$

discovered in 1905 by Albert Einstein. In natural units ($c = 1$), the formula expresses the identity

$$E = m \tag{1}$$

$$\int \oint \sum \prod \subset \supset \subseteq \supseteq \alpha\beta\gamma\rho\sigma\delta\epsilon \tag{2}$$

Let $\mathbf{u}$,$\mathbf{v}$ and $\mathbf{w}$ be three vectors in $\mathbf{R}^3$. The volume V of the parallelepiped with corners at the points $\mathbf{0}$, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{u}+\mathbf{v}$, $\mathbf{u}+\mathbf{w}$, $\mathbf{v}+\mathbf{w}$ and $\mathbf{u}+\mathbf{v}+\mathbf{w}$ is given by the formula

$$V = (\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}.$$

$$f(x, y, z) = 3y^2z \left(3 + \frac{7x + 5}{1 + y^2}\right) \tag{3}$$

In non-relativistic wave mechanics, the wave function $\psi(\mathbf{r}, t)$ of a particle satisfies the *Schrödinger Wave Equation*

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{-\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi + V\psi.$$

[1]Inside minipage

It is customary to normalize the wave equation by demanding that

$$\iiint_{\mathbf{R}^3} |\psi(\mathbf{r},0)|^2 \, dx\, dy\, dz = 1.$$

A simple calculation using the Schrödinger wave equation shows that

$$\frac{d}{dt}\iiint_{\mathbf{R}^3} |\psi(\mathbf{r},t)|^2 \, dx\, dy\, dz = 0,$$

and hence

$$\iiint_{\mathbf{R}^3} |\psi(\mathbf{r},t)|^2 \, dx\, dy\, dz = 1$$

for all times t. If we normalize the wave function in this 34 4 4.1 way then, for any (measurable) subset V of $\mathbf{R}^3$ and time t,

$$\iiint_{V} |\psi(\mathbf{r},t)|^2 \, dx\, dy\, dz$$

represents the probability that the particle is to be found within the region V at time t.

5 list

1. $d(x,y) \geq 0$ for all points x and y of X;
2. $d(x,y) = d(y,x)$ for all points x and y of X;
3. $d(x,z) \leq d(x,y) + d(y,z)$ for all points x, y and z of X;
4. $d(x,y) = 0$ if and only if the points x and y coincide.

- $d(x,y) \geq 0$ for all points x and y of X;
- $d(x,y) = d(y,x)$ for all points x and y of X;
- $d(x,z) \leq d(x,y) + d(y,z)$ for all points x, y and z of X;
- $d(x,y) = 0$ if and only if the points x and y coincide.

test1 AAAAAA

test2 AAAAAA

Figure 1: This is just a long figure caption for the minion in Despicable Me from Pixar

6 figure

7 box

This is an easy way to box text within a document!

8 box and code

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>

float **train_data;
int *lable;
int main()
{
    int row = 0;
    int col = 0;

    char buffer[10000];
    FILE *fp = fopen("../../NeuralNetC/iris.data", "rb");
    while (1) {
        if (!fgets(buffer, sizeof buffer, fp)) break;
        //printf("%s\n", buffer);
        char seps[] = ", \t\n\r";
        char *token;
        token = strtok(buffer, seps);
```

```
        int i = 0;
        while (token != NULL) {
            printf("%d, %s\n", i, token);
            if(i==0) row = atoi(token);
            if(i==1) col = atoi(token);
            token = strtok(NULL, seps);
            i++;
        }
        printf("%d %d", row, col);
    }
}

//float **train_data;
//int *lable;
//
//int main(int argc, const char * argv[]) {
//    freopen("../../NeuralNetC/iris.data", "r", stdin);
//    //freopen("test.out", "w", stdout);
//
//    int row;
//    int col;
//
//    scanf("%d", &row);
//    scanf("%d", &col);
//
//    printf("%d %d", row, col);
//
//    train_data = (float **) malloc (row * sizeof(float*));
//    lable = (int *) malloc (row * sizeof(int));
//    char label_str[255];
//    for (int i = 0; i < row; i++) {
//        train_data[i] = (float *) malloc ((col - 1) * sizeof(float));
//        for (int j = 0; j < col - 1; j++) {
//            scanf("%f", &train_data[i][j]);
//        }
//        scanf("%s",label_str);
//    }
//
//    for (int i = 0; i < row; i++) {
//        for (int j = 0; j < col - 1; j++) {
//            printf("%f ", train_data[i][j]);
//        }
//
//    }
//
//    fclose(stdin);
```

```
//    return 0;
//}
```

References

Ancey, C., Coussot, P., and Evesque, P. (1996). Examination of the possibility of a fluid-mechanics treatment of dense granular flows. *Mechanics of Cohesive-frictional Materials*, 1(4):385–403.

Radjavi, H. and Rosenthal, P. (1973). *Invariant Subspaces*. Springer-Verlag, New York.