RESEARCH ARTICLE OPEN ACCESS

Probability of Genetic Diagnosis

GanSiqing*, SunHeng**

*(School of Information Science, Jinan University, Guangzhou, Guangdong Province, China)

** School of Information Science, Jinan University, Guangzhou, Guangdong Province, China)

--************************----------------------------------

Abstract:

The main study in this paper is to calculate the probability of disease in patients with disease genes. At the time of gene sequence analysis, the distance between the DNA strand and the template strand was calculated using Levenshtein, and the probability of occurrence of the disease was estimated using similarity. Then we construct the DNA molecule on the probabilistic reasoning method for the Bayesian formula to be calculated. In order to test the validity of the sequences designed in this paper, nupack software was used to test and participate in the design of the sequence to detect the temperature of the reaction and the free energy of the secondary structure.

Keywords — **DNA Computing, Bayesian Formula, DNA Strand, The Levenshtein Distance.**

--************************----------------------------------

I. INTRODUCTION

According to the literature published by the scholars in the study of DNA computing and experiments, we can conclude that the following advantages of DNA computing[1]: 1) high parallelism. 2) large capacity storage capacity. 3) low energy consumption. 4) Resources[2-5].

The biomolecule calculation process requires a large number of molecules in the probability estimates, but the actual bioinverters should not be at the expense of the probability of calculating the number of molecules at the expense. Under the advantage of biomolecule computing, large-scale parallelism and complementary matching are the guiding subject. Therefore, this paper is based on DNA molecular fragments to calculate genetic diagnosis and probability calculation. The method reduces the number of steps and molecules. The DNA chains of similarity were calculated using existing knowledge such as the Levenshtein distance, and the probability of disease was calculated using similar probabilities. According to the characteristics of DNA fragment probability calculation, we can achieve Bayesian reasoning of DNA molecules.

In this paper, PSA disease as an example, based on molecular fragments of the probability calculation method. Levenshtein distance was used to improve the matching degree of single DNA calculation. Two single-stranded DNA matches were calculated based on Levenshtein distance, and the probability of DNA molecule was calculated by DNA strand displacement technique. To achieve two functions: 1, the known pathogenic factor gene and randomly selected genes to match, and according to its similarity to predict the degree of gene mutation .2, the probability of calculation in the DNA fragment.

II. PRINCIPLES OF THE MODEL

This paper adopts the Levenshtein distance to solve the problem of string matching, and calculate the similarity of two single-strand DNA,and calculate the similarity.On the distance of the string problems, the hamming distance method is mainly adopted, but this article adopted the method of

Levenshtein distance.The hamming distance is not very suitable for the DNA strand length which is more than 10 bp, and the DNA strand length is different.In this paper, the structure of any given prior probability, disease probability and signal probability of DNA structure, is inferred through the complementary DNA strand reaction posterior probability calculation, namely Bayesian calculation.

Levenshtein distance:In the 1960s V.Levenshtein first proposed the Levenshtein distance, refers to between two strings, the string x converted to string y with the required minimum number of edit operation $d_L(x,y)$.

Two string x,y edit distance related to both the public subsequence of the string.If: $x=a_1...a_m, y=b_1..b_n$ is the letters Σ on the two strings,and there is a string index $1\le i_1\le\cdots\le i_l\le m$ and $1\le j_1\le\cdots\le j_l\le n$ for all $k(1\le k\le l)$ have $\mathrm{c_k}=a_{i_k}=b_{j_k}$, then the string $z=c_1c_2\cdots c_l$ is x and y a common subsequence $l(x,y)$.

Obviously, an empty string is always a common subsequence. Therefore, any two strings common subsequence set are not empty. Between the string x and y the deletion of similarity is defined as the length of the longest common subsequence $l(x,y)$.

Edit distance is a measure of Σ^*.For Σ on any two strings x and y, satisfy:

$$d_l(x,y)=|x|+|y|-2l(x,y) \qquad (1)$$

Therefore,deleting the similarity is closely related to two strands and the maximum number of Watson-Cricket sequences matching. For two completely denying strand, the value happens to be their length.

similarity functions:Similarity function is used to measure single DNA molecular thermodynamic properties of similarity, It can be used as mathematical analysis carried out on the DNA strand. Assuming that Σ is an alphabet, Σ^* on the mapping of the similarity function is to satisfy the following conditions:

$$\sigma(x,x)\ge\sigma(x,y)=\sigma(y,x)\ge 0, x,y\in\Sigma^*$$

The similarity function corresponding to the Levenshtein distance is:

$$\sigma_\alpha(x,y)=n-d_l(x,y), x,y\in\Sigma^n$$

Each pair of string x and y in Σ^n will be mapped to the corresponding character position σ_α of the same position for two strings . This function turns to the Levenshtein similarity, and meets the condition:

$$\sigma_\alpha(x,y)=n-d_l(x,y), x,y\in\Sigma^n \qquad (2)$$

Random Variables: Possible values for the function is a random phenomenon.It can take different values of the field, so that we can talk about a continuous and discrete random variables or Boolean random variables.The article will focus on discrete variables, use the number of nucleotide strand of DNA as a probability.For example, we can talk about a random variable d on behalf of a given disease, meaning the probability for the D.

Logical propositions: Logic formula expresses its distribution between a random variable and the value in the potential field.

Probability function:In the random variable domain a function P distribution probability is assigned to each value (and therefore concludes from variables that every potential logical proposition). Building on an example, we can talk about the probability of D as the duple $P(D) =< P(D_1), P(D_0) >$ .The sum of probabilities of all the values of the domain must be equal to 3:

$$P(D_1) + P(D_0) = 1 \tag{3}$$

When we will define the function does not depend on other random variables, and we call it the prior probability.

Joint probability:Define a different set of propositions $a_1a_2......a_n$,they are joint probability function defined on the probability of the same time, expressed as $P(a_1 \wedge a_2 \wedge ... \wedge a_n)$ or $P(a_1, a_2,..., a_n)$.

Conditional probability:This function can be directly described as variable and other's trust after the first observation of relevant variables.So the conditional probability that the proposition a of a given b is the probability of a, known b.It can be derived from the conditional probability $P(a | b)$. Conditional probability can also be expressed as a function of prior probability and joint probability:

$$P(a | b) = \frac{P(a,b)}{P(b)} \tag{4}$$

This formula can be derived so-called product rule:

$$P(a,b) = P(a | b)P(b) = P(b | a)P(a) \tag{5}$$

When extensive research of disease, a disease d of probability given the signal s is known and expressed as $P(d | s)$. This is also called aposterior probability.

Conditional independence:Two propositions a and b are conditionally independent when they do not have any dependency relationship. In such case we can rewrite their probabilities as

$$P(a | b) = P(a);$$

$$P(b | a) = P(b);$$

$$P(a,\ b) = P(a)P(b) \tag{6}$$

Bayes' Law : Can be derived from the conditional probability and the product rule formulations, and is stated as follows:

$$P(d | s) = \frac{P(s | d)P(d)}{P(s)} \tag{7}$$

This rule, together with the property of independence, are key in probabilistic reasoning and allows the establishment of relationships between probabilities and evidences. It allows to update the certainty value of a hypothesis or a diagnosis (prior probability P(d)), in the light of new evidence (P(s)) and Probabilistic Reasoning with a Bayesian DNA Device the signal likelihood (P(s|d)), to obtain an "updated" posterior probability (P(d|s)).

III. GENETIC DIAGNOSIS

Genetic cloning of human prostate specific antigen (PSA) was isolated from a library of clay from GM607 lymphoid cell lines and PSA genes which have been sequenced completely.In the case of a given disease genes known PSA, we can calculate two strand matching degree and the similarity degree of two strands by the Levenshtein distance calculation.By calculating the probability of illness,we can carry on subsequent probabilistic reasoning.The normal gene with PSA length is different.So we can choose the Levenshtein distance to calculate the distance of the two strand, through public sequence of two strand probability calculation.

Process

Genetic diagnosis process is as follows:

Firstof all,we should extract no mutations from normal gene in diseased regio,flag compared to N.

Then, extract mutated gene in diseased region which is called PSA,flag compared to M.

And then,react normal genes with disease genes complementarily.

Finally, through observing the combined parts of a gene, the uncombined part of the combination is the conspicuous location in the disease genes,remember to C.

Using the knowledge of the theory of probability, we may safely draw the conclusion:

The probability of illness is $P = \frac{C}{M}$ (8).

DNA extraction

Because of the mutability and diversity of gene, gene in the process of growth is constantly changing,and some diseases are caused by a certain parts of the human body gene mutations.So the extraction of diseased region gene is the most correct sequence. Extracted DNA from genes is possible to be double-stranded DNA.So you need to denatured react double-stranded DNA to produce two single-strand DNA.The two single DNA markers for M and $\overline{M}$,As shown in figure 1 of the Commission.

As shown in table 2-1.A sequence of strand and strand a `, strand b and b ` is completely complementary, and between the strand and strand b is part of the complementary.In the strand of sequences a, b or complementary sequence of the same color in the same, with a dotted line the underlined sequences are not complementary pairing sequence in the reaction.

TABLE I

THE REACTION OF THE SEQUENCE

Reaction strand	The sequence of strand code
Strand a	5`-TAACACCCTCCGACC-3`
Strand a`	5`-GGTCGGAGGGTGTTA-3`
Strand b	5`-GGTGGATGAGAGTGTTA-3`
Strand b`	5`-TAACACTCTCATCCACC-3`

PSA partial gene fragment

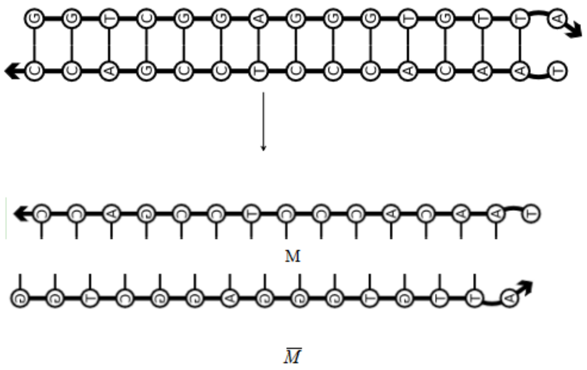


Fig.1. Double strand PSA genes into single-stranded DNA

To extract the gene treatment under test is the same as the figure 1.Through denaturedreaction a single-strand DNA could become double-stranded DNA, marked as N,as shown in figure 2.

Assume that genes under test N= 17

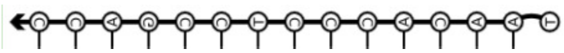


Fig. 2. test gene segment

Through the Watson - Cricket base pairing to PSA genes and genetic match reaction under test,as shown in figure 2.

Levenshtein Distance

Through the Watson - Cricket base pairing to PSA genes M=15 and genetic N= 17 match reaction under test,as shown in figure 3.

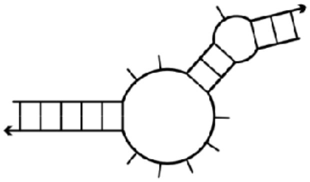


Fig.3. complementary reactive

According to the above assumptions M = 15, N = 17.The Levenshtein distance algorithm is used to calculate the two DNA single strand, and the Levenshtein distance is obtained:

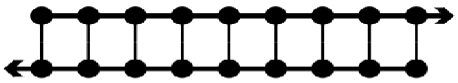
$$d_l(M,N) = |M| + |N| - 2l(M,N) = 15 + 17 - 2\times 9 = 32 - 18 = 14$$

For the corresponding similarity:

$$\sigma_\alpha(x,y) = n - d_l(x,y) = 32 - 14 = 18$$

Fig.4. two strands public sequence ($2\times l$)

The probability of illness:



$$p(M,N) = \frac{d_l(M,N)}{n} = \frac{18}{32}$$

IV. PROBABILISTIC REASONING

In order to realize the inference on the DNA molecule, we conduct the calculation according to the results obtained above.The Bayesian formula is calculated based on the calculation of molecular fragment, according to the total length of DNA as a sample space,and selecting a sample of DNA fragments.Assuming that the probability of illness is D, produced sick signals are S.There are two ways to get the signal S:Signal and no signal.Therefore the signal probability is 0.5.According to clinical experience we can draw a conclusion that sick signal probability is 0.7 on the premise of sick.Therefore,we can use DNA molecular to compute Bayesian probability.

$$P(\frac{D}{S}) = \frac{P(\frac{S}{D})\times P(D)}{P(S)} \quad (9)$$

According to the above:

$$p(M,N) = \frac{d_l(M,N)}{n} = \frac{18}{32}$$

Now make assumptions:

$$P(D) = \frac{d_l(M,N)}{n} = \frac{18}{32} \; P(S) = 0.5 \; P(\frac{S}{D}) = \frac{7}{10}$$

We will put these formula into common denominator (can not be directly calculated with points, but in order to get more clearer show that this paper will make a reduction to common fractional denominator) :

$$P(D) = \frac{45}{80} \; P(S) = \frac{40}{80} \quad P(\frac{S}{D}) = \frac{56}{80}$$

DNA molecule structure

When calculating the formula (9),We need to construct:

$$P(D)=\frac{45}{80}$$

$$P(S)=\frac{40}{80}$$

$P(\frac{S}{D})=\frac{56}{80}$.IAs the denominator is the same in the three probability, so we can directly make a sample space N=80.

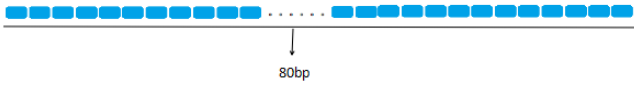


Fig. 5. sample space

Sample space represents the complete genome sequences.This article for each nucleotide sequence no actual meaning.Here to explain a new logic reasoning method of DNA molecular in detail, so the details will be skipped briefly.

In this paper, we have found out two public sequences of the gene and the common sequence is the DNA fragments of the probability $P(M,N)=P(D)=\frac{45}{80}$.But the probability calculation of the disease is obtained through two sequences with different length.In this section,for the convenience of calculation, we will unify the denominator of probability.The DNA fragments above also need to be changed.

1,Probability strand $P(\frac{S}{D})$ of DNA structure:There are two parts of single DNA complementary pairing.On the part of long segments of the DNA molecule hypothesis for the event $S \wedge D$,the following short segments of the DNA molecule is D.The DNA structure reacted by two single strand is the probability $P(\frac{S}{D})$.

2,Probability $P(D)$ of DNA structure:The probability $P(D)$s indicated by event D which is made up of single DNA molecule.

3,Probability $P(S)$ of DNA structure:probability $P(S)$ is similar with $P(D)$ probability.The DNA fragments made up of event S show the probability $P(S)$.

Inference procedure

In the diagnosis of the disease, we need to use the signal S to help diagnose disease D.In the case of the disease signals, the probability of disease can be measured directly.We according to the known data in the case of a given signal S to infer the disease incidence of the D.

According to the above calculation of prior probability of disease are as follows:

- P (D = presence) = 0.5625
- P (D = absent) = 0.4375

Hypothesis disease signal prior probability:

- P (S = presence) = 0.5
- P (S= absent) = o.5

According to the experimental data,the conditional probability of a disease signal is given:

- P (S = absentǀ D = absent) = 0.7
- P (S = presence ǀ D = absent) = 0.3
- P (S = absentǀ D = presence) = 0.2
- P (S = presence ǀ D = presence) = 0.8

We have confirmed the presence of disease signal (S =presence) , and the probability of disease also come out from the above calculation.Under the condition of a given disease, the disease signal probability P (D = = ǀ S exist) needs us to calculate.Although all of the data has been told,this article is reasoning operation of probability based on DNA fragments.So the Bayesian formula to calculate by hand directly and molecular computing Bayesian formula on the DNA molecule.Comparing the two results to evaluate whether it is right.

The data given in this directly using the Bayesian formula to calculate the results:

$$P(\frac{D}{S}) = P(\frac{S}{D}) \times P(D) \Big/ P(S) = \frac{56}{80} * \frac{45}{80} \Big/ \frac{40}{80} = \frac{63}{80} \tag{10}$$

In order to achieve the computing formula on the DNA molecule, we according to nucleotide construct DNA molecule, the corresponding probability of using DNA replacement method, the results of the calculation formula of 10.

ACKNOWLEDGMENT

In this article,compared with paper:Probabilistic Reasoning with a Bayesian DNA Device Based on Strand Displacement, we present more conditions.But we implement the new ideas and approaches presented in this paper, which greatly reduce the computational complexity of DNA computing.The DNA reasoning liberated from a large number of DNA molecules.

Thesis writing is boring hard and challenging.Calculation logic of the DNA molecule is now the front end of the project of science and technology.With great enthusiasm to this project, write this article.Here, I especially want to thank my mentor Heng Sun teacher.Without HengSun teacher's hard cultivation, pursuit of teaching, there is no the smooth completion of my thesis.

In the future work will be dedicated to the research on the DNA molecule logic reasoning, with the knowledge of probability theory summary and generalization.

REFERENCES

[1] Adleman L: Molecular computation of solutions to combinatorial problems.Science 1994, 266:1021-1024.

[2] Lipton, R.J.: DNA solution of hard computational problems. Science 268(5210),542﹣545 (1995).

[3] Adar, R., Benenson, Y., Linshiz, G., Rosner, A., Tishby, N., Shapiro, E.: Stochastic computing with biomolecular automata. Proceedings of the National Academy of Sciences of the United States of America 101(27), 9960﹣9965 (2004).

[4] Benenson, Y., Adar, R., Paz-Elizur, T., Livneh, Z., Shapiro, E.: DNA molecule provides a computing machine with both data and fuel. Proc. Natl. Acad. Sci.USA 100(5), 2191﹣2196 (2003).

[5] Benenson, Y., Gil, B., Ben-Dor, U., Adar, R., Shapiro, E.: An autonomous molecular computer for logical control of gene expression. Nature 429, 423﹣429 (2004).