

Probability Computation of Molecular Matrices

SiqingGan¹, Heng Sun²

¹ School of Information and Science Technology, Jinan University,
Guangzhou, 510632, China, E-mail: 2460246481@qq.com

² School of Information and Science Technology, Jinan University,
Guangzhou, 510632, China, E-mail: tsunheng@jnu.edu.cn.

(Received July 29, 2017, accepted September 13, 2017)

Abstract. We propose an automatic, programmable and computational model consisting of biomolecules[1-5] by transforming a base pair into a one-dimensional matrix containing only 0,1, using the matrix length as the sample space, representing the event The sample point takes the percentage of the sample space as the probability of the event, and details the probability calculation problem into the model of the molecular calculation problem. After the introduction of the molecular matrix calculation probability method, the examples are given to illustrate the realization of the complex event molecular matrix calculation probability. In order to verify the feasibility and complexity of calculating the probability problem of the molecular matrix, we use the manual calculation probability to compare with the results of the DNA chain to predict the DNA secondary structure and its interaction through NUPACK[6] software.

Keywords: biomolecular equipment, probability calculation, Bayesian probability problem, matrix.

1. Introduction

In 1959, Feynman first proposed the concept of calculation at the molecular level[7], but subject to the level of materials and equipment at the time, molecular computing and did not get too much attention. At the same time, in the field of biological research, molecular biology is gradually emerging and mature, biology research into the molecular level. In the 1980s, as the understanding of molecular biology theory deepened and the modern biochemistry and bioengineering technology became more and more perfect, the material basis for molecular calculation was basically available. In order to achieve a comprehensive computer reform and computer bottlenecks breakthrough, biological computing by the computer field experts and scholars to introduce.

In 1994, Professor Leonard Adleman of the University of Southern California used the base (Deoxyribonucleic Acid, DNA) molecule as the calculation medium, relying on modern molecular biotechnology as a means to successfully solve the seven vertex directional Hamiltonian paths in the biological laboratory Problem (Hamiltonian Path Problem, HPP)^[8]. After Leonard Adleman's groundbreaking realization of biomolecule calculations, various methods and procedures have been proposed. Biomolecular calculations were first used to solve the problem of NP-complete problems^[9], and gradually developed into nanotechnology and biomedical-oriented applications such as genetic diagnostics and drug delivery automata[10-13].

In the reasoning flow of DNA molecules, the characteristics of DNA molecular computing are also vaguely displayed. DNA calculation of the following advantages: 1) high degree of concurrency[14]; 2) mass storage capacity; 3) low power consumption; 4) resource rich; 5) Watson-Crick matching ability. DNA chain in the calculation of mathematical problems such as NP complete problem, you can achieve multi-location real-time concurrent computing, the number of its number can be calculated, the current estimated DNA computing power can be more than 10⁵ times the supercomputer, significant savings in computing time. In DNA storage, 1nm DNA molecular solution can store 1 billion billion of binary data, mass storage capacity to achieve only a limited space to calculate a large amount of data stored. DNA energy used in the calculation of molecular energy, only the enzyme catalysis and heating operation can be completed, according to statistics, the energy consumed by the computer is the computer to complete the same calculation of the energy consumed by one billionth of a billion molecules Can be completed 10¹⁹ times the parallel computing, and the traditional computer consumption of the same energy can only complete 10⁹ operations. DNA computing resources and energy required are green resources and controllable, so the energy consumption will not be harmful to nature, DNA calculation of raw materials are

deoxyribonucleotides and restriction enzymes, can be recycled. DNA can be found in almost all organisms, synthetic DNA chain technology has also become mature, so the DNA used to calculate the raw materials are resource-rich.

The matrix uses the values of rows and columns as indexes to find the position within the matrix and manipulate the data for that position. The calculation of the matrix is suitable for the operation of multidimensional arrays and the storage and operation of large data, mainly by obtaining the values of the rows and columns of the matrix. The two-dimensional matrix will effectively locate the data in the plane, easy to operate, combined with the DNA chain structure characteristics, this paper will store the DNA chain information in the form of two-dimensional array. It is easy for us to think of the coherence of the mapping of the DNA matrix, using the DNA chain as the sample space, the number of deoxyribonucleotides as the number of events, Watson-Crick matching the DNA strand as an operation, to achieve a series of probability calculations.

DNA chain corresponding to the base pair x , $x \in \{A, C, G, T\}$, $A = T, C \equiv G$, you can assume that A and T correspond to the number '1', C and G correspond to the number '0'. In the matrix consisting only of '0' and '1', the base of the DNA strand corresponds to the numbers '0' and '1' in the matrix. The directivity of the DNA sequence can be represented by the position of the matrix, and the length of the DNA sequence is represented by Matrix size representation. The sample space S is represented by a matrix in which the values in the matrix represent the sample points, the size of the matrix represents the number of sample spaces, and the matrix assignment is '0' at initialization. When event A occurs, the sample point of event A in matrix is converted from '0' to '1' by number '0', and the number of '1' can indicate that there is sample number of event A. In calculating the probability of occurrence of event A, we first calculate the transpose matrix of the event A matrix, that is, the antisense strand of the DNA strand, and the probability value is the number of A and T in the DNA double strand divided by the total number. The use of double-stranded DNA to calculate the probability greatly reduces the DNA strand in the calculation of the error, to ensure that the data can not be modified, in the calculation of the probability of DNA double-stranded to match the situation, compared to the DNA single-stranded base easier. This problem of molecular probability calculation is transformed into a matrix calculation problem.

The number of deoxyribonucleotides (SS) was determined by PCR amplification as the sample space and event. The number of deoxyribonucleotides was used as the corresponding sample space and the number of events. The DNA was obtained by DNA Watson-Crick Matching and limiting the enzyme to the DNA strand to identify the site to operate, the cohesive end attracting recognition site complementary, stimulate the restriction endonuclease response, the specified location of the cut to achieve the calculation of the probability of the process, the event molecules attached to fluorescent molecules and quenching Agent, after the operation, showing a different fluorescence, the observer observed by laser irradiation fluorescence color calculation results.

2. Algorithm Example Submitting

The algorithm of calculating the probability calculation of the molecular matrix needs to ensure the reliability and security of the algorithm, and the probability problem is calculated accurately at the lowest time and space. The realization of the algorithm not only to be able to meet the probability of a single calculation, as well as to achieve the universality of the calculation template, the similarity probability calculation can also be completed quickly. In the realization of strong confidentiality at the same time to ensure the simplicity of the algorithm to verify the convenience of the algorithm.

DNA calculation steps:

Step 1: Construct the sample space. To understand the purpose of the test, estimate the sample number n of the sample space S . In this paper, we use the form of a molecular matrix to represent $S = [0 \ 0 \ 0 \ \dots \ 0 \ 0 \ \dots \ 0 \ 0]$, $\text{length}(S) = n$, $x_i = S[i]$, $i = 1, 2, \dots, n$. The DNA strand

$S = 5' - x_1 x_2 x_3 \dots x_{i-1} x_i \dots x_{n-1} x_n - 3'$, $n \in N^+$, $x_i \in \{A, T\}$, $i = 1, 2, \dots, n$ required for the synthesis of the sample space S using n bases.

Step 2: Construct the event. Write the number according to the test to the sample point, assuming the meaning of the event. Use the molecular matrix to represent the sample of the event, such as event $A = [x_1 \ x_2 \ \dots \ x_{i-1} \ x_i \ \dots \ x_{n-1} \ x_n]$, $\det(A) = \sum_{i=1}^n x_i$, $\text{length}(A) = n$, $x_i \in \{0, 1\}$, where x_i represents the i -th sample point