ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 4, 2017, *pp*.304-311

Probability Computation of Molecular Matrices

SiqingGan[1], Heng Sun [2]
[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 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 Leomard 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 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 The 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 ... 0 0 ... 0 0], $\text{length}(S)=n$, $x_i = S[i]$, $i=1,2,...,n.$ The DNA strand $S = 5`-x_1x_2x_3...x_{i-1}x_i...x_{n-1}x_n-3`$, $n \in N^+, x_i \in \{A,T\}, i=1,2,...,n$ 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 = [x1 x2 ... xi-1 xi ... xn-1 xn] , $\det(A) = \sum_{i=1}^{n} x_i$, length (A) = n, $x_i \in \{0,1\}$, where xi represents the i-th sample point

In the case of xi = 1, the event contains the sample point of the i-bit, xi = 0 indicates that the event does not contain the sample point of the i-bit, and det (A) indicates that the event contains the number of sample points. Molecular chain with n base synthesis event DNA strand $A = 5`-x_1x_2x_3...x_{i-1}x_i...x_{n-1}x_n-3`,\ n \in N^+, x_i \in \{A,C,G,T\}, i = 1,2,...,n$.

Step three: operation. Deal with events and events between the operation, get new events. According to the molecular matrix, the molecular matrix of the corresponding event is calculated by the concept of the relationship between events and events.

Step 4: Calculate the event probability. The probability of the event calculation The operation in the re-molecular matrix is the number of samples obtained by multiplying the matrix and the matrix's permutation matrix. The ratio of the number of samples to the sample space is the probability of the event. The probability of seeking the event in the molecule is to release the base, and the single-stranded event is matched to the double-stranded base, and the probability is obtained according to the ratio of the base A, T and base A, C, G, T consumed at the time of matching.

Step five: repeat steps three, four, remove all the error solution, get the final result.

Pseudo code description:

Input:int S； int A[S]， B[S]；

Output：p//The probability of the event.

Fake code：//To achieve the process of calculating the probability.

```
int S//The sample space S is an integer constant.
int A[S]//Event A is a one-dimensional array, the length of the array is the number of sample space, the array contains only 0 and 1.
int B[S]//Event B is a one-dimensional array, the length of the array for the sample space, the array contains only 0 and 1.
int length//The new sample space is used to mark the corresponding sample space for the new event.
while(true)
    int C[length]//Create an array to store new events.
    for(int i=0;i<S;i++)
        C[i]=match(A[i],B[i])//The result of the operation between the events is stored in the new event.
        Length=match(A[i],B[i]).length;//The corresponding sample space for the new event is used to store the sample space after the new event has passed.
    End for
    if(match(C[length])==true)
        int det=0;//Defines and initializes the number of samples for the event.
        for(int i=0;i<length(C);i++)
            det=det+C[i];//The number of samples contained in the calculation event, that is, the number of events in the event.
        End for
        break;//Get the required array: End the loop.
    End if
End while
int p=det(C)/length(C);//The probability of the event output.
End
```

In the algorithm, the time complexity is O (n2) and the spatial complexity is O (n). The algorithm is simple and easy to understand. The algorithm uses A and T bases to denote 1, C and G represent 0, and the molecular calculation contains four kinds of ACGT Base, in the crack algorithm is required to calculate 4n, to achieve the security of the algorithm.

3. Probability Instance

Conditional probability: Let A, B be two events and are called conditional probability of occurrence of condition A in event B.Remember：

$$P(A \mid B) = \frac{P(AB)}{P(B)} \tag{1}$$

Example: What is the probability of being divisible by 3 in the case of an even number in an integer 1-10?
Solution: Initialize the sample space S = [0 0 0 0 0 0 0 0 0 0], length (S) = 10.

Let A be a number that can be divisible by 2, A = [0 1 0 1 0 1 0 1 0 1], det (A) = 5, length (A) = 10

Event B is the number that can be divisible by 3, B = [0 0 1 0 0 1 0 0 1 0], det (B) = 3, length (B) = 10

Event C is a number that can be divisible by 2 and divided by 3, C = [0 0 0 0 0 0 0 0 0 0], det (C) = 1, length (C) = 10

Event D is the number that can be divisible by 3 when it is known to be divisible by 2, D = [0 0 1 0 0], det (D) = 1, length (D) = 5

Table 4-1 Molecular chain representation of events

Event	Molecular chain	Antisense molecular chain	Number of samples	Number of sample spaces
A	5`-CACTCACTCA-3`	3`-TCACTCACTC-5`	5	10
B	5`-CCACCTCCAC-3`	3`-CTCCACCTCC-5`	3	10
C	5`-CCCCCACCCC-3`	3`-CCCCCTCCCC-5`	1	10
D	5`-CCACC-3`	3`-CCTCC-5`	1	5

Table 4-1 shows the molecular chain representation of the event, the red represents the base of the matrix element as '1', the black represents the base of the matrix of '0', and the purple represents the sample point of the event. From event A, event B, and event C, we can see that event C is the product of event A and event B. From event matrix A, event B, and event D, we can see that event D is a conditional event for event B under the condition of known event A.

At the temperature of 27.0 C, the molecular chain in the table was analyzed by nupack software. The structure and basic indexes of the molecule were shown in the following figure.

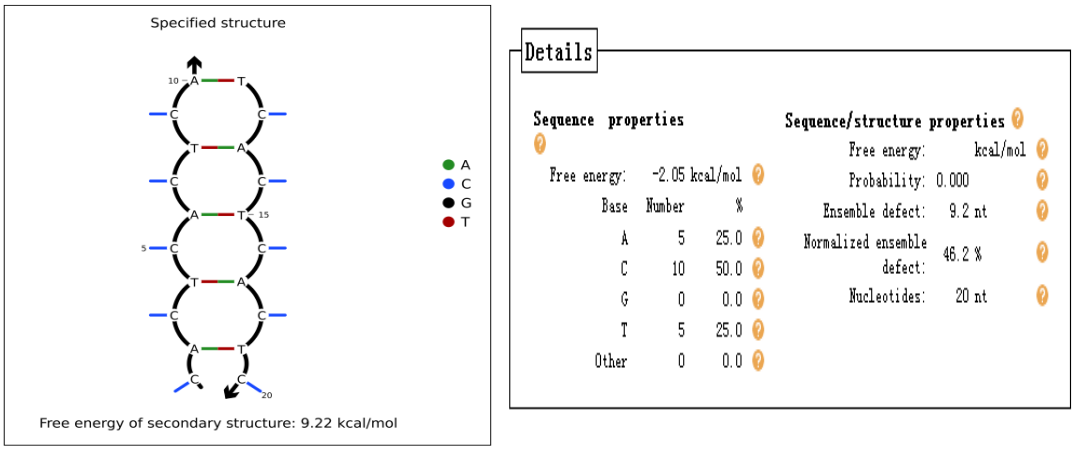
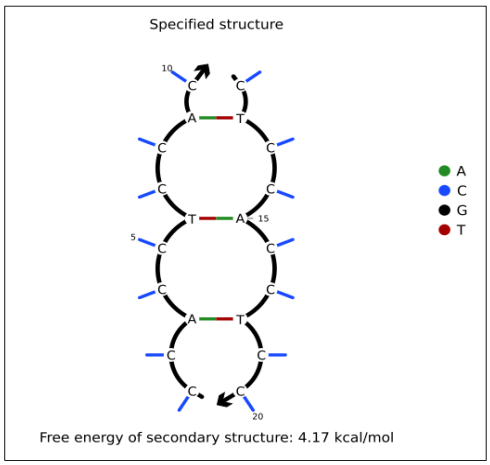


Figure 4-1 Event A

The probability of event A is: $P(A) = \frac{\det(A)}{length(A)} = \frac{5}{10}$

Event A probability of the molecular matrix calculation:

$$P(A)=\begin{bmatrix}0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1\end{bmatrix}\begin{bmatrix}0\\1\\0\\1\\0\\1\\0\\1\\0\\1\end{bmatrix}=\frac{5}{10}$$



Details

Sequence properties

Free energy: -1.38 kcal/mol

Base	Number	%
A	3	15.0
C	14	70.0
G	0	0.0
T	3	15.0
Other	0	0.0

Sequence/structure properties

Free energy: kcal/mol
Probability: 0.000
Ensemble defect: 4.8 nt
Normalized ensemble defect: 23.8 %
Nucleotides: 20 nt

Figure 4-2 Event B

The probability of event B is: $P(B)=\frac{\det(B)}{length(B)}=\frac{3}{10}$

Event B probability of the molecular matrix calculation:

$$P(B)=[0\quad 0\quad 1\quad 0\quad 0\quad 1\quad 0\quad 0\quad 1\quad 0]\begin{bmatrix}0\\0\\1\\0\\0\\1\\0\\0\\1\\0\end{bmatrix}=\frac{3}{10}$$

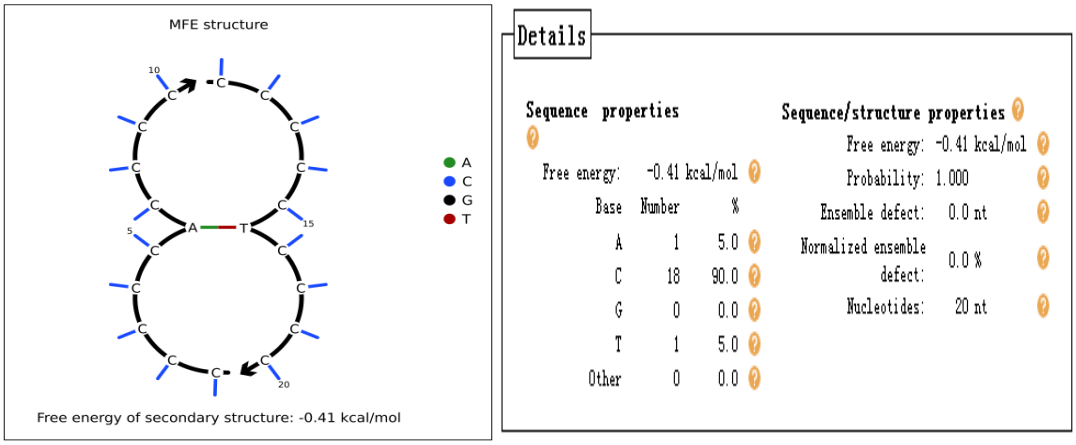


Figure 4-3 Event C

The probability of event C is: $\mathrm{P(C)}=\frac{\det(C)}{length(C)}=\frac{1}{10}$

Event C probability of the molecular matrix calculation:

$$P(C)=\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix}=\frac{1}{10}$$

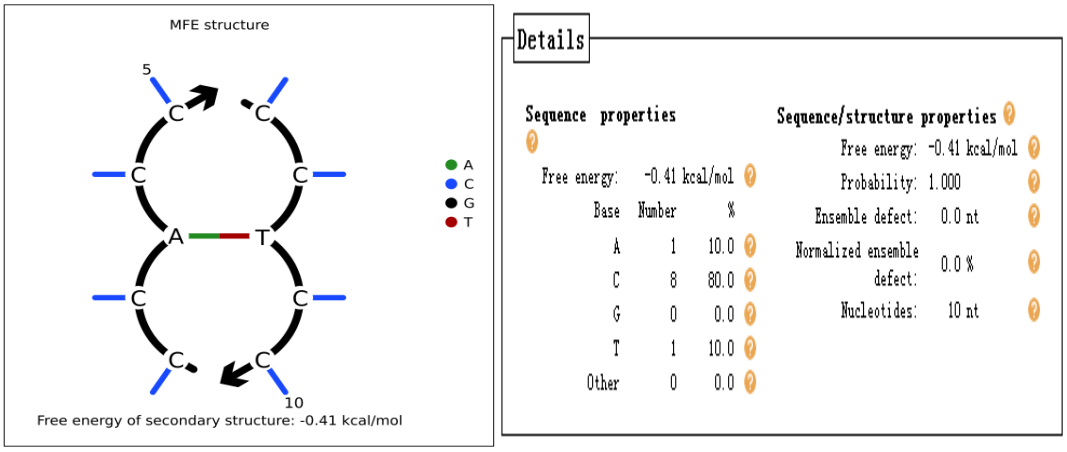


Figure 4-4 Event D

The probability of event D is: $P(D)=\frac{\det(D)}{length(D)}=\frac{1}{5}$

Event D probability of the molecular matrix calculation:

$$P(D)=\begin{bmatrix}0 & 0 & 1 & 0 & 0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}=\frac{1}{5}$$

Calculations of the molecular matrix can be seen in the calculation of the difference event probability, the first to calculate the event A and event B of the event C, probability P (C) = P (AB). And then the conditional event D is calculated by the molecular matrix A, B, C and D. The sample space in the event event D is the sample number of the event A, the sample point of the event D is the sample point of the event C corresponding to the sample point in A (6) corresponds to D [2] = 0, A [6] corresponds to D [3] = 0, and A [4] corresponds to D [1] = 0, A [4] A [8] corresponds to D [4] = 0, A [10] corresponds to D [5] = 0, and C [5] = A [5], so D [3] = 1. The probability of the difference event is obtained by the ratio of the total number of bases A and T consumed by the molecular chain forming the double strand in the event of event D, which accounts for the total number of bases consumed. P (D) = P (C) / P (A) can be derived from the derivation of the above molecular matrix.

4. Summary and Prospect

Molecular matrix method to achieve the basic events and complex events of the probability of calculation have achieved good results, experimental consumption of materials in a relatively small number of cases to start a large number of data calculation, the process easy to understand, the experiment is simple and accurate calculation. In the test, according to the data for simulation, due to economic and experimental conditions, and did not use the gene for the actual operation, only through the genetic software for testing.

Outlook: Although this paper only carries out theoretical reasoning, the actual operation has yet to be verified, but we can use this method to clarify the feasibility and practicality of this method. From the construction and derivation of this paper, it can be seen that it is simple and feasible to carry out the actual probability calculation on the numerator, such as the matrix in the probability theory, so we can conclude that as long as we find the law of development, we can Phenomenon to see the essence. How do DNA molecules maintain their life activities in life, and now that they can not find the law, but believe that in the near future, humans can solve the mystery and use them for the benefit of the cause of mankind. The same theory of DNA molecular reasoning and modern known theory have the same point, as long as we can carefully observe the nature, to be analyzed, to find a common law, we can find deep law.

5. References

[1]. Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H. & Weiss, R. A synthetic multicellular system for programmed pattern formation. Nature 434, 1130–1134 (2005).
[2]. Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. Science 314, 1585–1588 (2006).
[3]. Rinaudo, K. et al. A universal RNAi-based logic evaluator that operates in mammalian cells. Nature Biotechnol. 25, 795–801 (2007).
[4]. Deans, T. L., Cantor, C. R. & Collins, J. J. A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells. Cell 130, 363–372 (2007).
[5]. Kobayashi, H.etal.Programmablecells:interfacing natural and engineeredgene networks. Proc. Natl Acad. Sci. USA 101, 8414–8419 (2004).
[6]. Dirks R. M. , Bois J. S. , Schaeffer J. M. , Winfree E. , Pierce N. A. , SIAM Rev. , 2007, 49, 65—88
[7]. Feynman RP: There's plenty of room at the bottom. Miniaturization.1961:282-296
[8]. Adleman L: Molecular computation of solutions to combinatorial problems.Science 1994, 266:1021-1024
[9]. Lipton, R.J.: DNA solution of hard computational problems. Science 268(5210),542–545 (1995)
[10]. 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)
[11]. 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)
[12]. 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)
[13]. Benenson, Y., Paz-Elizur, T., Adar, R., Keinan, E., Livneh, Z., Shapiro, E.:Programmable and autonomous computing machine made of biomolecules. Nature 414(6862), 430–434 (2001)
[14]. Feynman bottom．ThechnicalAnnual plenty report，The oftheAmerican Institute of Meeting Physical Society, Califomia Technology, December959 29，1 SeemanNC．Nucleic－acidandlattices．JournalofTheoretical