GMDH - GA Hybrid Model Extracting Exon Region from DNA Sequences

Kouji OHTA[1], Ikuo YOSHIHARA[2], Kunihito YAMAMORI[3], Moritoshi YASUNAGA[4]

Abstract

A model building method based on Group Method of Data Handling (GMDH) optimized by GA is developed for extracting exon regions. GMDH, that is originally a method to construct higher order polynomial model, is extended to constructing higher order logical model.

The model built by proposed method is compared with Genetic Programming (GP)-based model as to the extraction rate of best, worst and average. The proposed method is superior to GP as to extraction rate of all for intron-exon boundary.

Key Word :

GMDH, Genetic Algorithm, DNA Sequences, Exon Region

1. Introduction

The genome information is embedded in DNA sequences which consist of four kinds of bases; A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). Human genome consists of 3 billion base pairs. Exons are protein-coding regions and occupy about 5% in all base sequence of human genome and other parts of DNA sequences are useless intron.

There are enormous data of genome sequences, so, it is very important to extract exon regions automatically and efficiently. The first two bases of intron are almost 'GT' and the last two bases are almost 'AG'. Only this boundary is used this time. However, they exist not only on the boundary but also inside the exons and introns.

Large amount of research have been performed for extracting exon region, for example, Bayesian Estimation, Neural Network (NN) and so on are widely used for extracting exon region. However, there are a few problems in researches. For example, In case of NN, it is difficult to design the number of layer and node properly. In order to avoid the matter, we develop a model building method based on Group Method of Data Handling (GMDH) optimized by Genetic Algorithm (GA) for extracting exon region. GMDH is a method to build a nonlinear model. The user has to decide selection and combination of explanation variables when building the model of GMDH. It is a very difficult process. We solve these problems and optimize model structure of GMDH using GA. We make experiments to extract the exon-intron boundary and the intron-exon boundary by proposed method, and compare proposed method with GP of other method.

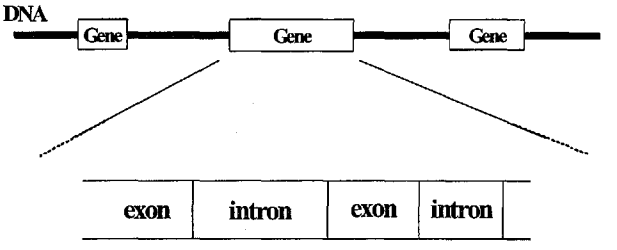

Fig.1 DNA Structure

2. Optimization of GMDH-Based Model by GA

GMDH is a method to build a model of non-liner system. We propose a method to decide a structure of GMDH model by GA.

2.1. GMDH

GMDH repeats processes which combine a transfer function, and builds a model as shown in Fig.2. Therefore, GMDH combines a simple quadratic model for built a complex nonlinear model. The transfer function of GMDH generally used as follows.

$$G(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2 \quad (1)$$

The parameters "a_0, $a_{1,\ldots}$, a_5" of each transfer function are determined using the least square method.

1) Postgraduate student, Graduate School of Engineering

2) Professor, Dept. of Computer Science and Systems Engineering.

3) Associate Professor, Dept. of Computer Science and Systems Engineering.

4) Professor, Inst of Information and Electronics, University of Tsukuba.

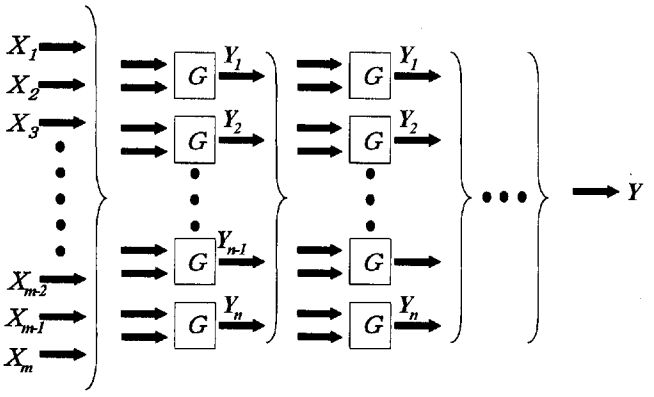

Fig.2 A Model Structure of GMDH

2.2. Extension of GMDH

The model of GMDH is expressed by multitiered structure. The data use in this research is binary data. Therefore, the transfer function of model for used logical operators (Table.1). For this reason, the model structure of GMDH can be expressed as a tree structure. Nodes are transfer function and leaves are explanation variables. We experiments by limiting a tree size for simplification of problem. Moreover, we cut the number of nodes in half every combination progress.

Table.1 Logical operators

Logical operator
AND
OR
XOR
NAND
NOR
EQV

We presume that the following formula is same meaning as equation (1)

$$G(x_1,x_2)=a_1(x_1\cdot x_2)+a_2(x_1+x_2)+a_3(x_1\oplus x_2) \quad (2)$$

(a_i: not or nop)

2.3. GA

2.3.1. Initial Population Generation

As initial population, individual structure models that have the explanation variable and the node function at random are generated.

2.3.2. Genetic Coding

A gene expresses two strings. The string shows the input of an explanation variable and another string shows the logical operator into each node. In the case of extracting an exon-intron boundary, the built model outputs 1 (True) for boundary or 0 (False) for a non-boundary. The dummy variable "d" is introduced into the explanation variable portion. The dummy variable "d" is the meaning chosen nothing.

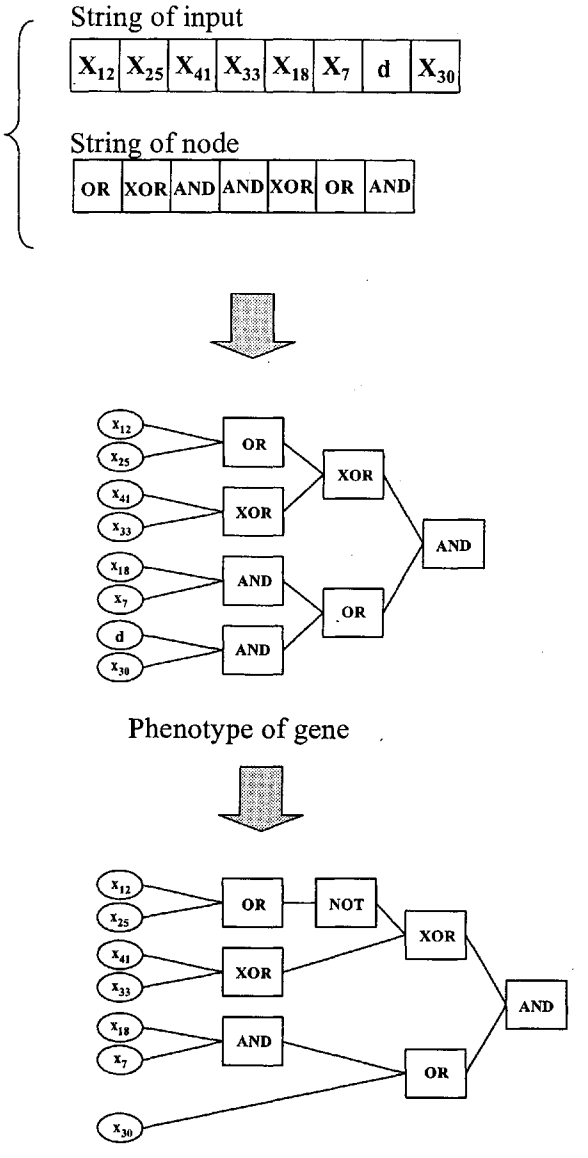

Fig.3 Example of Genetic Coding

2.3.3. Crossover

As crossover method, we use uniform crossover. Two individuals are selected at random from the population as a pair of parents. We made bit sequence of a mask pattern at random, and parent's gene is inherited according to a mask pattern. The fig.4 shows the example of uniform crossover to produce two children.

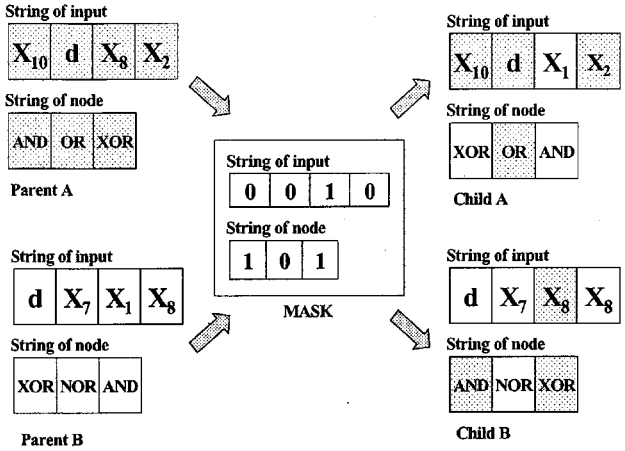

Fig.4 Example of Crossover

2.3.4. The Fitness Evaluation

Each individual are estimated by the fitness. The fitness is a measure and shows the next generation the probability of surviving. The fitness is computed from the following formulas.

$$fitness = \sum_{i=1}^{t} b_i + \sum_{j=1}^{u} n_j \qquad (3)$$

(t : Boundary data for built model, u : Non-boundary data for built model, b : The number of correct answers in boundary data, n : The number of correct answers in non-boundary data)

In the data of built model, the fitness is total of the number of correct answer.

2.3.5. Natural Selection

Ranking selection is employed for natural selection. Individuals are ranked depending on fitness, and chosen from the highest fitness up to the population size.

2.3.6. Mutation

If a mutation happens, the explanation variable changes to other explanation variable at random depending on mutation rate. However, mutation is not performed to the elite.

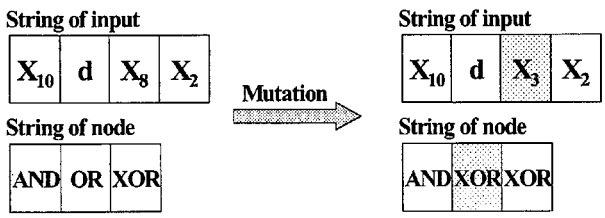

Fig.5 Example of Mutation

3. Verification Experiments for Extracting Exon Region

We experiment to verify the effectiveness of the proposed method. The GP was taken up as a candidate for comparison and we perform a comparison experiment of extracting Exon Region by proposed method.

3.1. Base Sequence Data

Large amounts of human genome data are released on the web site of NCBI (National Center for Biotechnology Information), from where we can obtain experimental data. This site shows if 'GT's is boundary or not.

The first two bases of intron are 'GT' and the last two bases are 'AG'. The boundary data of exon-intron consists of total 22-base from upstream 10-base and downstream 10-base. The boundary data of intron-exon consists of total 39-base from upstream 30-base and downstream 7-base.

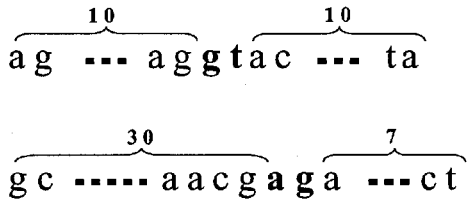

Fig.6 A piece of base sequence data

Four kinds of bases are coded to binary as follows.
A = (1, 1), G = (0, 1), C = (1, 0), T = (1, 1)
Fig.7 shows an example of binary coding of base sequences. Input data is 44 bit string which coding depending on the above rule. In case of AG, input data is 78 bit string.

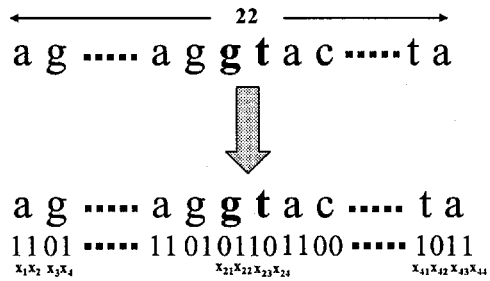

Fig.7 binary coding of base sequence data

3.2. The Measure of Extraction Rate

We define two amounts of statistics for evaluation of model reliability. One is called sensitivity (Sn) and divides the number of exons predicted correctly by the actual number of exons. The other is called specificity (Sp) and divides the number of exons predicted correctly by the number of exons predicted. In fact, Sn is sensitivity and mean percentage of correct calculation

among the true exon. Sp is specificity and mean percentage of correct calculation for predicted exon.

$$Sensitivity = \frac{b}{B} \times 100 \quad (\%) \qquad (3)$$

$$Specificity = \frac{b}{b+n'} \times 100 \quad (\%) \qquad (4)$$

In our research, we define Sensitivity' (Sn') to evaluate the identification rate of non-boundary.

$$Sensitivity' = \frac{n}{N} \times 100 \quad (\%) \qquad (5)$$

identified	Boundary (b)	Non-boundary (b')	Boundary (n)	Non-boundary (n')
actual	Boundary (B)		Non-boundary (N)	

All data

Fig.8 The result of extraction

3.3. Experimental Conditions

Experimental conditions are as follows.

<Genetic algorithm>

- Population size 100
- Crossover rates 100%
- Mutation rates 0.01%
- Maximum generation 3000

<Base sequences data>

Exon-intron boundary (GT boundary)
- Boundary 2100
- Non-boundary 20000
- Total number of data 22100

Intron-exon boundary (AG boundary)
- Boundary 2100
- Non-boundary 20000
- Total number of data 22100

The detail of base sequences data
- Model building 2000
- Verification 20100

The number of data we choose at random for model building is 1000 from boundary data and 1000 from non-boundary data, and other data are used for verification.

3.4. Experimental Results

Fig.9 and Fig.10 shows experimental results using the proposed method and GP.

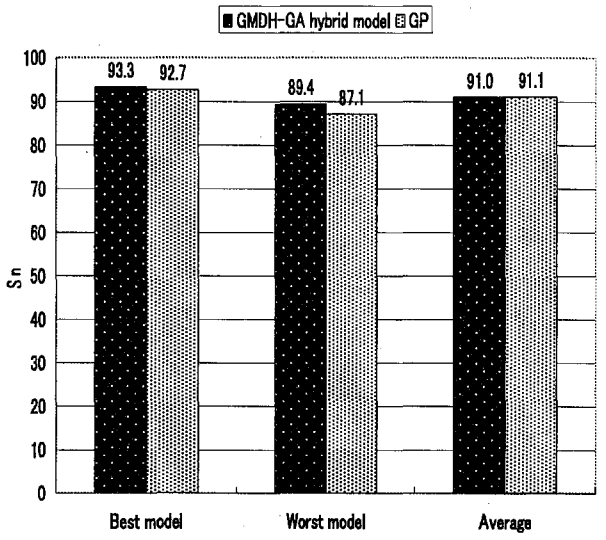

Fig.9 The extraction rate of GT boundary of Sn

The average extraction rate by the proposed method is 91.0% for Sn. The proposed method is superior to GP as to the extraction rate of best model and worst model. But, the average is slightly inferior to GP.

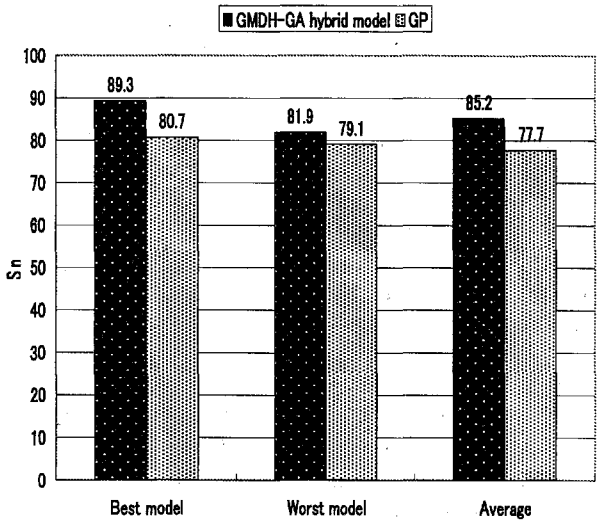

Fig.10 The extraction rate of AG boundary of Sn

The average extraction rate by using proposed method is 85.2% for Sn. The proposed method is superior to GP as to all kinds of extraction rate of Sn.

3.5. Frequency of Appearance of Explanation Variable

Frequency of appearance of the explanation variable is shown in Fig.11 and Fig.12 as input of best model.

Frequency of appearance of near GT boundary and near 'AG' boundary is shown in Fig.13 and Fig.14. A numerical value higher than 0.2 in Fig.13 and Fig.14 shows that specific base tends to appear. We compare

frequency of appearance of the base contained in a model with correlation of a base of near boundary. Therefore, it is thought that the proposed method can extract the important factor automatically.

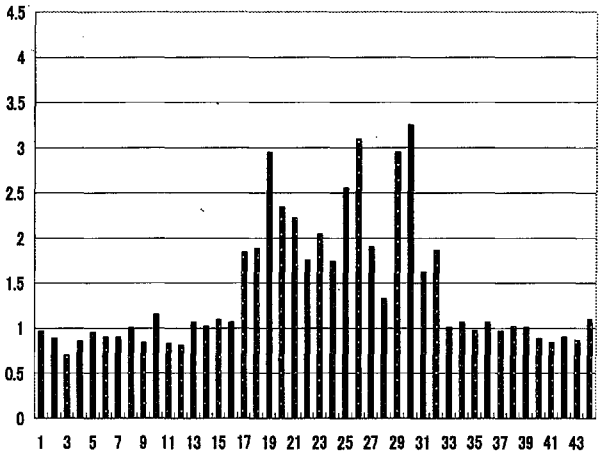

Fig.11 Frequency of appearance of the explanation variable near GT boundary

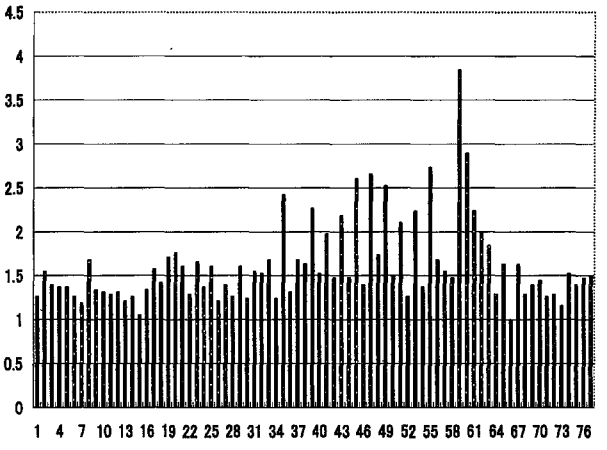

Fig.12 Frequency of appearance of the explanation variable near AG boundary

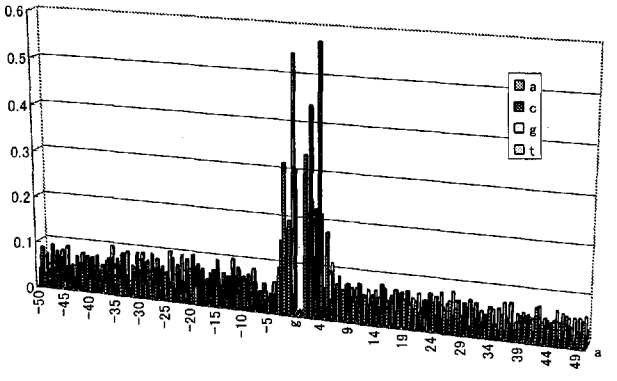

Fig.13 Frequency of appearance near GT boundary

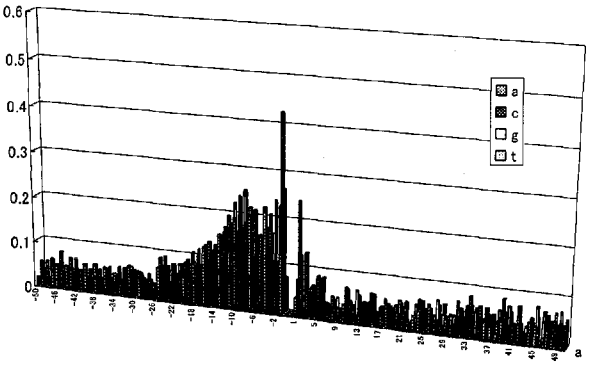

Fig.14 Frequency of appearance near AG boundary

4. Conclusion

The proposed method builds a model for extraction automatically and selects optimal explanation variable. We compare proposed method and GP by extraction rate of boundary. In case of proposed method, average extraction rate of 'GT' is 91.0% for Sn. The proposed method is the same level as the extraction rate of GP. Average extraction rate of 'AG' is 85.2% for Sn. The proposed method is superior to GP as to all kinds of extraction rate of Sn. Extraction of AG boundary is much more difficult than extraction of GT boundary. Therefore, the proposed method is superior to GP.
Moreover; we compare frequency of appearance of the base contained in a model with correlation of a base of near boundary. Therefore, it is thought that the proposed method can extract the important factor automatically.

Future works will be to improve extraction rate of 'GT' and 'AG'.

5. Acknowledgements

This research is partly supported by MEXT grant 16011204 in 2004.

6. References

[1] Kanehisa, M., Invitation to Genome Information, kyoritsu shuppan Co., ltd., 1996, (in Japanese)

[2] Muramatu, M., Genome 2, medical science international Co., 2000,(in Japanese)

[3] Kitano, H., Genetic Algorithm, sangyo tosho publishing Co., ltd., 1993 (in Japanese)

[4] Mitaku, S., Kanehisa, M., A Human Genome Project And Knowledge Information Processing, baifukan Co., ltd., 1995 (in Japanese)

[5] Ikeda S., "The Foundations and Application of GMDH", systems and control, Vol.23, No.12, pp.710-717,(1979)

[6] Ikeda S., "The Foundations and Application of GMDH", systems and control, Vol.24, No.1, pp.46-54,(1980) ,(in Japanese)

[7] Yoshihara I., Sato S., "Nonlinear model Building Method with GA and GMDH", Information Processing Society of Japan. ICS, Vol. 96 Num. 78 pp.1-6 (1996.08), (in Japanese)

[8] Mitaku, S., Sakaki, Y., Bioinformatics, tokyo kagaku dozin Co.,ltd., 2003 (in Japanese)

[9] Goldberg, D. E., GENETIC ALGORITHMS in Search, Optimization, and Machine Learning, Addison-Wesley, Inc, 1989

[10] National Center for Biotechnology Information <URL>http://www.ncbi.nlm.nih.gov/