Development of Exon Region Extracting Method by GMDH and GA 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 models, is extended to constructing complex logical model.

The proposed method automatically builds a model for extraction and selects optimal important parts of DNA sequence..

Key Word :
GMDH, Genetic Algorithm, DNA Sequences, Exon Region

1. Introduction

Human genome consists of 3 billion base pairs. It is divided into several kinds of regions, for example exon, intron, etc. Exons are the protein-coding DNA sequences, but introns are not. There are enormous data of genome sequences, so, it is very important to extract exon regions automatically and efficiently.

Various kinds of researches have been performed for exon extraction, for example, Bayesian Estimation, Neural Network (NN) and so on. However, there are a few problems in researches. For example, regarding NN, it is difficult to design the number of layer and node properly. In order to avoid the matter, we propose 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 researchers have to decide selection and combination of explanation variables when building the model of GMDH. It is a very difficult process. We intend to solve these problems and optimize model structure of GMDH using GA.

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 sciences and Electronics, Univ. of Tsukuba

We proposed GMDH-based model optimized by GA for extracting exon regions from DNA sequences, and make experiments to extract GT boundaries and AG boundaries in DNA sequences of non-human genome.

2. Genome Information

2.1. Genome

Genome information's are embedded in DNA sequences which consist of four kinds of bases; A (Adenine), C (Cytosine), G (Guanine), T (thymine). Human genome consists of 3 billion base pairs. Exons are protein-coding regions and occupy about 1.5% in all base sequence of human genome. Fig.1 shows the mechanism of protein compound from DNA sequences.

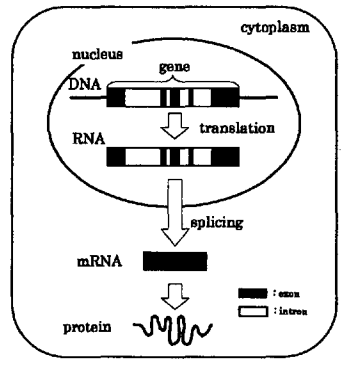

Fig 1 The mechanism of protein compound

2.2. Exon boundary and intron boundary

Boundaries between exon and intron have a feature. The fiest two bases of introns are almost GT and the last two bases are almost AG (fig.2). Only this boundary is used this time. But extracting the boundary between exon and intron is difficult. They exist not only on the boundary but also inside the exons and introns. We develop the method for extracting the boundary between exon and intron.

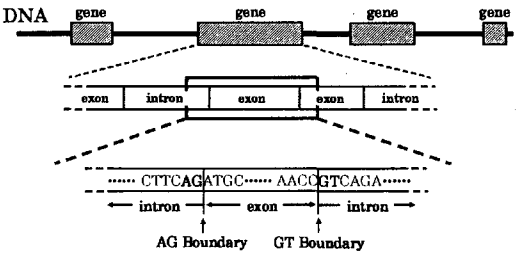

Fig.2 GT boundary and AG boundary

2.3. Base sequence data

We set base sequence data in consideration of the result of preliminary experiments as follows. The boundary data of exon-intron consists of total 22-base string from upstream 10-base string and downstream 10-base string. The boundary data of intron-exon consists of total 39-base string from upstream 30-base string and downstream 7-base string. 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.

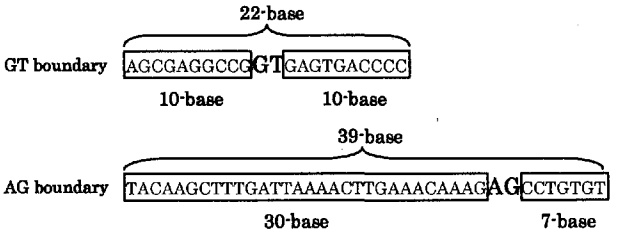

Fig 3 A Piece of Base Sequence Data

3. Optimizing GMDH-Based model by GA

3.1. GMDH

GMDH repeats processes which combine a transfer functions, and builds a model as shown in Fig.2. Therefore, GMDH combines a simple quadratic model for building a complex nonlinear model. The transfer function of GMDH generally is used eq.1. The parameters "a_0, a_1,..., a_5" of each transfer function are determined by using the least square method.

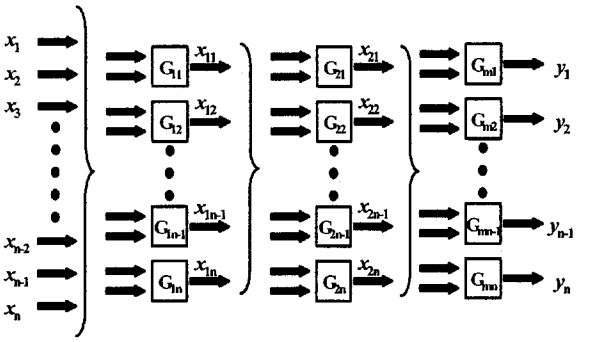

Fig.4 A model structure of GMDH

$$G_{mn}(x_i, x_j) = a_{mn}^{(0)} + a_{mn}^{(1)} x_i + a_{mn}^{(2)} x_j + a_{mn}^{(3)} x_i x_j + a_{mn}^{(4)} x_i^2 + a_{mn}^{(5)} x_j^2 \quad (1)$$

3.2. Extension of GMDH to logical function

Four kinds of bases are represented by binary string (1: True, 0: False) as follows.

A = (1, 1), G = (0, 1), C = (0, 0), T = (1, 0). (2)

Fig.5 shows an example of binary coding of base sequences. Input data are a 44-bit string which is coded according to the above-mentioned rule. As regards AG, input data is a 78-bit string.

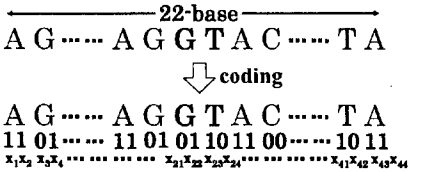

Fig.5 Binary coding of base sequence data

The transfer function of GMDH generally is polynomial. Therefore, eq.1 is extended to eq.3. 'op' is

logical operator and the true table of logical operator used for this research is shown table.1.

$$G(x_i, x_j) = x_i \; op \; x_j$$
$$op = \begin{bmatrix} AND & NAND \\ OR & NOR \\ XOR & EQV \end{bmatrix} \quad (3)$$

Table.1 true table

Input		Output $G(x_i, x_j)$					
x_i	x_j	AND	OR	XOR	NAND	NOR	EQV
0	0	0	0	0	1	1	1
0	1	0	1	1	1	0	0
1	0	0	1	1	1	0	0
1	1	1	1	0	0	0	1

3.3. GMDH model optimized by GA

The researchers have to decide selection and combination of explanation variables when building the model of GMDH. It is a very difficult process. We intend to solve these problems and optimize model structure of GMDH using GA.

3.3.1. GA and genetic operation

The procedure of GMDH model building using GA is shown in Fig.6.

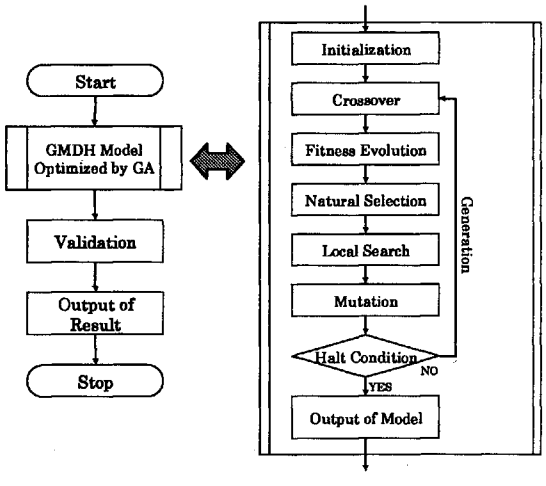

Fig.6 Flow chart of GMDH model building by GA

4. Validation experiments for extraction exon region

4.1. The measure of extraction rate

We define two measures 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 means percentage of correct calculation for predicted exon.

$$Sensitivity = \frac{b}{B} \times 100 \quad (\%) \quad (4)$$

$$Specificity = \frac{b}{b + n'} \times 100 \quad (\%) \quad (5)$$

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

$$Sensitivity' = \frac{n}{N} \times 100 \quad (\%) \quad (6)$$

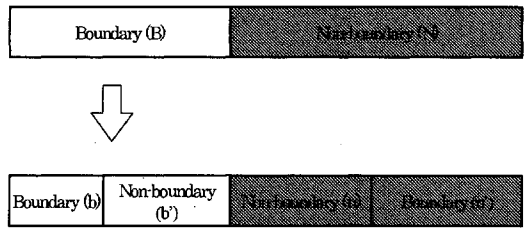

Fig.7 The Result of Extraction

4.2. Experimental conditions

Experimental conditions are as follows, and the number of data is shown in table.2.

<Genetic algorithm>

- Population size 100
- Crossover rates 100%
- Mutation rates 0.01%
- Maximum generation 4000

Table.2 The number of data

	GT boundary			AG boundary		
	Boundary	Non-boundary	Total	Boundary	Non-boundary	Total
The number of data	2092	19864	21956	2006	17011	19017

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 validation.

4.3. Experimental results

Table.3 show experimental results using the proposed method.

Table.3 Extraction rate by GMDH

		Extraction rate (%)		
		Average	Best model	Worst model
GT	Boundary (Sn)	90.6	94.2	86.4
	Non-boundary (Sn')	85.2	89.8	80.4
AG	Boundary (Sn)	86.1	91.0	79.7
	Non-boundary (Sn')	78.5	83.1	73.1

Table.2 shows extraction rate. As for exon-intron boundary, the average extraction rate is 90.6% for boundary and 85.2% for non-boundary. As for intron-exon boundary, the average extraction rate is 86.1% for boundary and 78.5% for non-boundary.

4.4. Frequency of appearance of explanation variable

Frequency of appearance of the explanation variable is shown in Fig.8 and Fig.9 as input of best model

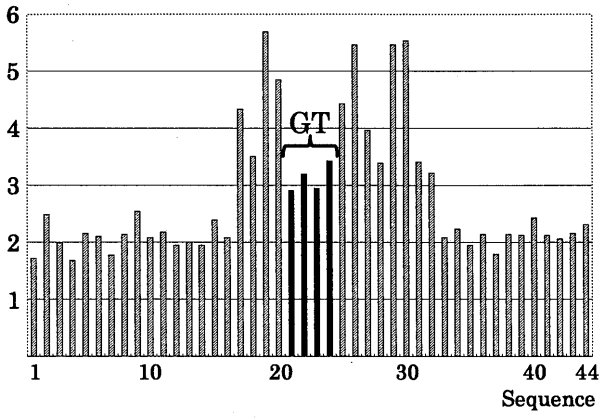

Fig.8 Frequency of Appearance of The Explanation Variable near GT Boundary

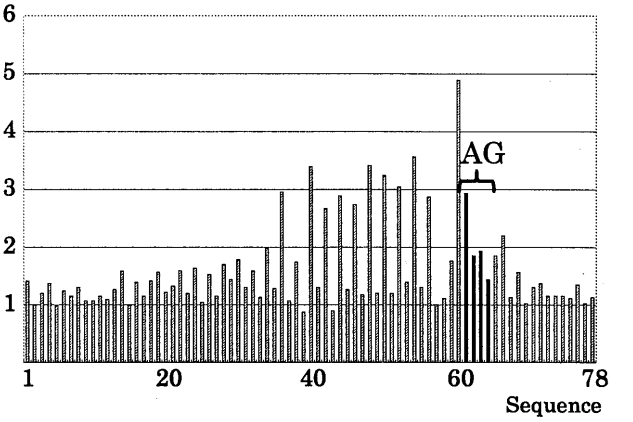

Fig.9 Frequency of Appearance of The Explanation Variable near AG Boundary

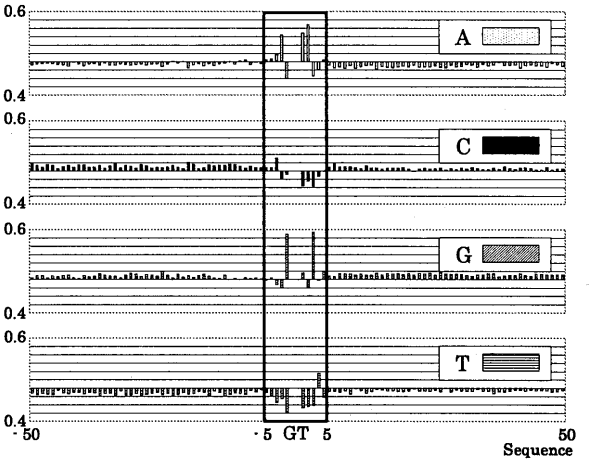

Fig.10 Frequency of Appearance near GT Boundary

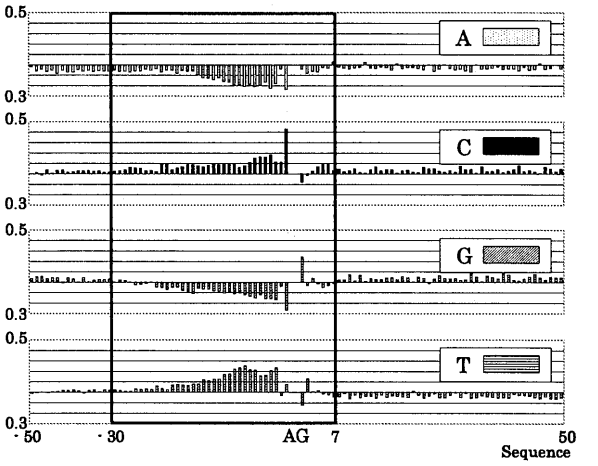

Fig.11 Frequency of Appearance near AG Boundary

Frequency of appearance near GT boundary and near AG boundary is shown in Fig.10 and Fig.11, respectively. A numerical value in Fig.10 and Fig.11 shows that specific bases often appear. We compare frequency of appearance

of the base contained in a model with frequency of appearance of bases near boundary. It is thought that the proposed method can extract the feature pattern of distribution automatically.

4.5. Application to other living creatures

We make experiments to extract GT boundaries and AG boundaries in DNA sequences of other living creatures. The living creatures are as follows.

Hs...Homo sapiens
Mm...Mus musculus
Dm...Drosophila melanogaster
Ce...Caenorhabditis elegans
Pf...Plasmodium falciparum
At...Arabidopsis thaliana
Sp...Schizosaccharomyces pombe
Sc...Saccharomyces cerevisiae

These experiment data were supplied by Univ. of Miyazaki, and we use the GMDH model of preceding section in validation. The number of data is shown in table.4.

Table.4 The number of data of living creatures

	GT boundary		AG boundary	
	Boundary	Non-boundary	Boundary	Non-boundary
Hs	366	22188	367	25594
Mm	352	15898	353	17308
Dm	195	7369	196	7452
Ce	176	3140	181	4306
Pf	87	2941	98	4024
At	626	18435	630	17096
Sp	75	5611	85	4078
Sc	58	5939	90	7085

Fig.12-15 show experimental results using the proposed method. Fig.12 and Fig.13 shows in GT boundary, and Fig.14 and Fig.15 shows in AG boundary.

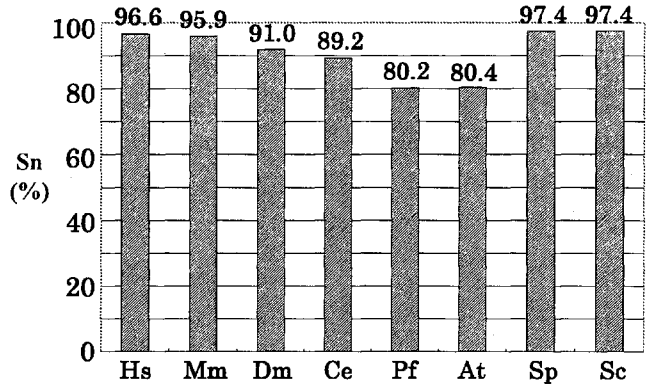

Fig.12 The Extraction Rate of GT Boundary of Sn

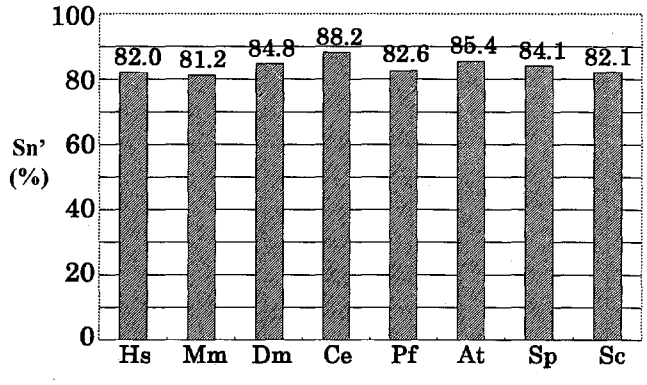

Fig.13 The Extraction Rate of GT Boundary of Sn'

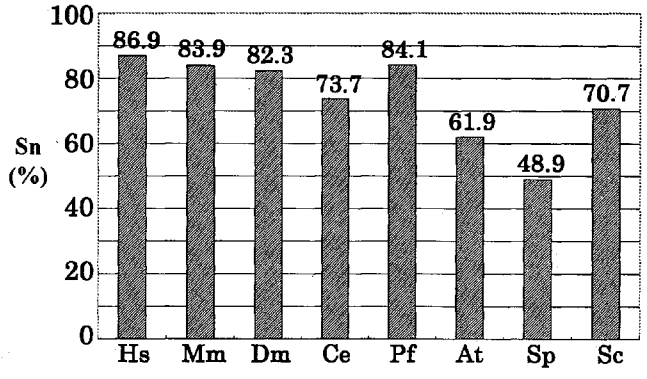

Fig.14 The Extraction Rate of AG Boundary of Sn

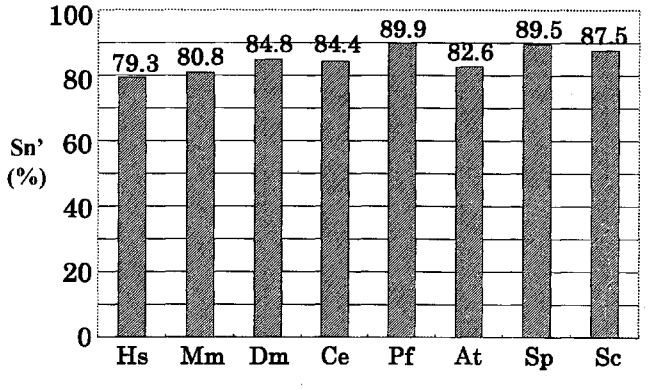

Fig.15 The Extraction Rate of AG Boundary of Sn'

As for Sn of GT boundary, the extraction rates of Mm, Dm, Ce, Pf, and At are lower than human (Hs), and Sp and Sc are higher. As for Sn of AG boundary, the extraction rates of non-humans creatures are much lower than human. I thought that there is much difference of extraction rates of living creatures, but there is not much difference of it. As for Sn' of GT boundary and AG boundary, the extraction rates of each non-human creatures are almost same. It is thought that exon region has unique DNA sequences of living creature.

5. Conclusion

We proposed GMDH-based model optimized by GA for extracting exon regions from DNA sequences. The proposed method automatically builds a model for extraction and selects optimal explanation variable. As for GT boundary, average extraction rate is 90.6% for boundary and 85.2% for non-boundary. As for AG boundary, average extraction rate is 86.1% for boundary and 78.5% for non-boundary. Moreover, we compare frequency of appearance of the bases contained in a model with that of near boundary and discuss the correlation. It is thought that the proposed method can extract the feature pattern of distribution automatically.

We make experiments to extract GT boundaries and AG boundaries in DNA sequences of non-human creatures. I thought that there is much difference of extraction rates of living creatures, but there is not much difference of it. As for Sn' of GT boundary and AG boundary, the extraction rates of each non-human creatures are almost same. It is thought that exon region has unique DNA sequences of living creature.

6. Acknowledgements

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

References

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

[2] Mitaku S, Sakaki Y, Bioinformatics, Tokyo Kagaku Dozin Co.,ltd., 2003 (in Japanese)

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

[4] John S. and Dan Roth: "Gene recognition based on DAG shortest paths", BIOINFORMATICS, Oxford University Press, pp.s56-s64, 2001

[5] A.G.Ivakhnenko, "The Group Method of Data Handling, A Rival of the Method of Stochastic Approximation", Soviet Automatic Control, Vol.13 No.3, pp.43-55 (1968)

[6] A.G.Ivakhnenko, "Polynomial theory of complex systems", IEEE Trans. on Systems, Man, and Cybernetics, Vol.SMC-1, No.4, pp.364-378.

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

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

[9] Yoshihara I, Sato S, "Nonlinear model Building Method with GA and GMDH", Information Processing Society of Japan. ICS, Vol. 96 pp.1-6 (1996) (in Japanese)

[10] Goldberg DE., GENETIC ALGORITHMS in Search, Optimization, and Machine Learning, Addison-Wesley, Inc, 1989

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

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