http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (4): 480-483, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.480.483

Research on DNA Encoding Based on Rough Set

[1]Song Yuan and [2]Zhengyao Hu
[1]College of Power and Mechanical Engineering, Wuhan University, Wuhan, 430072, China
[2]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, China

Abstract: One of the most fundamental issues in DNA computing is how to improve the Signal to Noise Ratio through effective encoding in the process of DNA computing. This paper propose a new DNA encoding filter selection algorithm based on Rough Set and Fuzzy theory: reduction the encoding standards expected through the Rough Set theory and filter the DNA encoding sets using the negative filter selection algorithm based on Rough Set, gain the initial encoding sequence library and construct an optimized method for DNA encoding selection algorithm based on Rough Set theory. Experimental results indicate that the method can effectively weaken the complexity of DNA encoding caused by the invalid constraint conditions, provide a new tool for the DNA encoding selection algorithm.

Key words: Rough set, DNA encoding, reduction, clustering, filter selection

INTRODUCTION

Richard Feynman first proposed the idea of molecular computing in 1961 (Xu and Zhang, 2008). Adleman published an article which solved a seven-vertex directed Hamiltonian Path Problem through specific biochemical tests for the first time in the U.S. journal Science in 1994 (Zhu *et al*., 2006). DNA computing has become a hot direction in the field of computer research in recent years for its high degree of parallelism, storage capacity, low power consumption, etc, as the emergence of bottlenecks in chip manufacturing technology and the popularity of parallel and distributed computing.

DNA computing divided into three stages: encoding, calculation produce solution space and the extraction solution (Yin *et al*., 2010). The aim of encoding is encoded each information element of the DNA to make it to be uniquely identified maximally in the actual biochemical reaction process. One of the traditional selection algorithms of DNA encoding is the search algorithm and the other is the construction algorithm. Because so much classification of the expected criteria in the evaluation of DNA encoding, the typical case is to choose the combination of several constraint conditions. However, due to the differences and overlaps of various constraints, a number of constraints are not as important as the others, or even some constraint are redundant, the results of the evaluation will be quite different when choosing a different combination of constraint conditions, the error estimation is larger and not comprehensive enough. In fact the constraint conditions which affect the DNA sequence to be identified uniquely may be the dominant little of all constraint conditions, too many constraints can lead to the increase of the noise and increase the complexity of the algorithm in choosing a combination constraint and some irrelevant constraint conditions will have an interference effect.

This study, based on previous studies, in view of the fuzziness and the classification rules which contained in the sequence similarity, the sequence stability and the overlapping and ambiguity of constraints in the process of DNA encoding, by introducing the Rough Set theory to reduction the encoding standards, then gain the initial encoding sequence sets through reverse selection of the basic encoding sequence library using clustering algorithm for mixed attributes based on rough set, constructed an optimized method for DNA encoding selection algorithm using Rough Set theory, The results show that the method can effectively weaken the complexity of DNA encoding caused by the invalid constraints and can improve the traditional DNA encoding selection algorithm.

DNA ENCODING BASED ON ROUGH SET

Rough set theory (Banerjee *et al*., 2007) is a theory of data analysis which proposed by the Polish mathematician Z. Pawlak in 1982 which is a new mathematical tool dealing with fuzzy and uncertain knowledge (Momin *et al*., 2006). The rough set theory is understanding the knowledge as the division of data

Corresponding Author: Zhengyao Hu, Wuhan University of Science and Technology, China

(Jensen and Shen, 2007), approximate description the imprecise and uncertain knowledge under the precondition of keep the same classification ability using the existing knowledge base, export decision and classification rules through knowledge reduction (Yang and Yang, 2008; Parthalain and Shen, 2010).

DNA encoding expected standards reduction algorithm: Building knowledge systems I = ⟨U, Ω, V, f⟩ based on Rough Set, U represents the DNA encoding sequence set, Ω content all kinds of constraint condition R and the similarity degree of DNA sequence S, V is the numerical description of DNA encoding sequence satisfy the constraint condition, f description the relationship between the attribute value and the numerical with language, firstly classify DNA sequences according to the extent of DNA encoding sequences satisfy the various constraint conditions and then delete the irrelevant or unimportant constraints conditions through the knowledge reduction, gain a subset of constraints to form the ultimate combination of constraints. The step of reduction algorithm as follows:

- Calculate the similarity of DNA encoding sequence, get the similarity matrix St
- Classify the DNA encoding sequence according to St, get the equivalence classes U/S
- Calculate the numerical description M of the DNA encoding sequence satisfy the constraint condition by the corresponding formula according to the set R
- Classify and sort by constraint conditions with M, disperse the numerical process, constitute set V
- Gain the subset of the constraint conditions R_1, $R_1 \subseteq R$ through Rough Set reduction rules

The key step in the process of the algorithm is the disposal of data discretization in order to get an equivalence class, the nature of it is a data clustering problem (Yao and Zhao, 2009) and this paper uses the Rough-based enhanced K-means clustering algorithm which provided by Fan and Wang (2010). The main process of the algorithm as follows: randomly select an object as initial cluster centers firstly, then new objects are classified by introduce two variables LT and UT, generate k cluster centers with self-learning algorithm, finally, the object is assigned to the upper and lower approximation set the most similar clusters according to the rules and update the cluster center, repeat until the criterion function for convergence is met. Take the calculation of equivalence class S for example, using the DNA sequences provided by (Zhang *et al*., 2006), as shown in Table 1. Based on the results provided, obtained similarity matrix St:

Table 1: DNA sequence

No.	DNA sequence	Length
A	ATGGTGCACCTGA..... GCCCTGGGCAG	92
B	ATGCTGACTGCTG..... GCCCTGGGCAG	86
C	ATGGTGCACTTGA.....GCCCTTGGCAG	92
D	ATGGTGCACTGGA.....GCCCTGGCCAG	92
E	ATGACTTTGCTGA.....GCCTTGGGCAG	92
F	ATGCTTCCACCTG.....CCCTGGGCACG	94
G	ATGGTGCATCTGT.....AGGCCCTGCGC	90
H	ATGCTGCACCTAA.....GCCCTGGGCAG	92

$$St = \begin{pmatrix} 0.0000 & 3.0670 & 3.7498 & 4.0642 & 1.7581 & 1.3849 & 2.0272 & 2.9775 \\ 3.0670 & 0.0000 & 3.6290 & 4.0714 & 3.8714 & 3.3102 & 2.4482 & 3.1812 \\ 3.7498 & 3.6290 & 0.0000 & 2.7780 & 4.1866 & 3.3184 & 3.6278 & 2.5889 \\ 4.0642 & 4.0714 & 2.7780 & 0.0000 & 4.4929 & 3.3487 & 3.8690 & 1.7753 \\ 1.7581 & 3.8714 & 4.1866 & 4.4929 & 0.0000 & 1.9167 & 2.2565 & 3.7554 \\ 1.3849 & 3.3102 & 3.3184 & 3.3487 & 1.9167 & 0.0000 & 2.4243 & 2.5025 \\ 2.0272 & 2.4482 & 3.6278 & 3.8690 & 2.2565 & 2.4243 & 0.0000 & 3.1842 \\ 2.9775 & 3.1812 & 2.5889 & 1.7753 & 3.7554 & 2.5025 & 3.1842 & 0.0000 \end{pmatrix}$$

Based on the Rough-based enhanced K-means clustering algorithm [10], take LT = 1.8, UT = 2.0 get the classification result: U/S = {{A, E, F}, {H, D}, {G}, {C}, {B}}.

Negative filter selection algorithm: This algorithm deal based encoding space with a fast reverse filtration refers to the numerical description M of the DNA encoding sequence satisfies the subset of the constraint conditions R_1, M and R_1 have been gained at the expected standards Reduction Algorithm. The step as follows:

- Screening the part of matrix M which matching R_1, composing the matrix M_R
- Sort matrix M_R based on the meaning of the numerical value each line, from matrix MS_R
- Select the front part of the DNA sequence refers to MS_R each line
- Choose the results that constitute the initial encoding sequence

Reduction algorithm based on rough set: After gain the subset of the constraint conditions R_1, the next step is filter and select the DNA encoding, this study introduces the Rough Set Theory, first, deal the based encoding space with a fast reverse filtration based on R_1, finally, gain the target encoding space based on R_1, the overall framework of the algorithm shown in Fig. 1, the step as follows:

- Input DNA encoding sequence, establish the basic encoding space U

- Select the constraint conditions; gain the subset of the constraint conditions R_1 through the reduction algorithm
- Gain the initial sequence library through the negative filter selection algorithm
- Search the initial sequence database based on R_1
- Get the target encoding set

RESULTS

Download the DNA sequence data from the NCBI database (http://www.ncbi.nlm.nih.gov), experiment the proposed method. Select sequences with the length of 85 to 95 randomly, in order to verify its feasibility, the DNA sequence S = {A, B, C, D, E, F, G, H} of Table 1 is labeled and added to the test data, filter out the overlap. Initial conditions contents the Hamming distance, the maximum length of the same sequences, H-measure, continuity constraints, hairpin constraints, reverse constraints, reverse-complement, melting temperature, GC content, free energy. Experiment with the RSES (Rough Set Exploration System) tool which integrates a variety of Rough Set-related functions, including discretization, attribute reduction, etc. The result showed in Table 2 with the number of Base DNA sequence 237, 536 and 781, the comparison of the practical iteration number when the algorithm complexity is O (n^4) shown in Fig. 2.

Experimental results show that the method is feasible and effective refer to the similarity results of DNA sequence in Table 1. Compared with the traditional DNA sequence encoding algorithms, this paper introduces the Rough Set theory, avoided the judgments and choices of constraint conditions depend on experience of the traditional algorithm through the constraint condition reduction based on Rough Set Theory, simplified the construction process of traditional search algorithm and reduced its complexity too. The proposed method can also be applied to improve the traditional algorithm as shown in Fig. 1, combined with the initial encoding sequence and the target space, can optimize their search algorithm by reducing the constraint condition of their search strategy.

Table 2: Filter results

No. of base sequence	No. of initial Sequence	No. of target sequence	Tag filter results
237	166	110	B, C, D, E, G
536	467	309	A, B, C, D, E, G
781	649	507	A, B, C, E, G, H

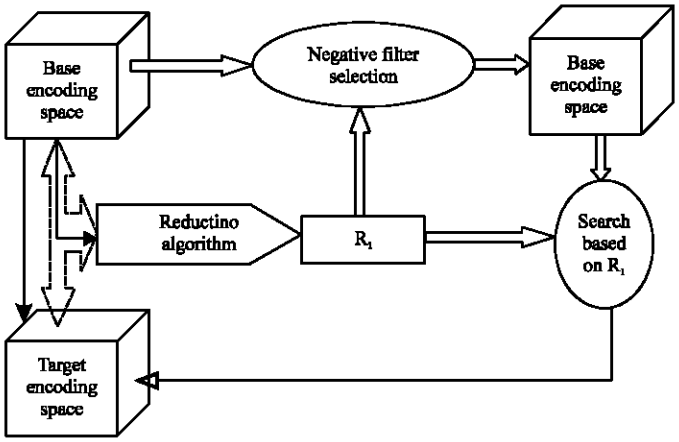

Fig. 1: Reduction selection algorithm based on rough set

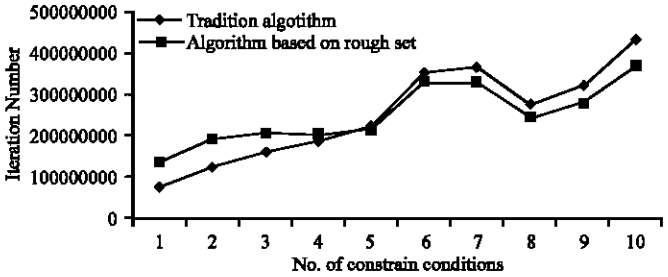

Fig. 2: Comparison of algorithm complexity

CONCLUSION

Aiming at the question of the uncertainty of the choice and judgments of constraint conditions depend on the experience of the traditional DNA encoding construction algorithm, this paper proposes an DNA encoding filtering selection algorithm according to the fuzzy constraint condition based on DNA encoding expected standards and its boundary by introduced Rough Set Theory, experimental results show that this method is feasible and effective. This method reduced and weakens the constraint of constraint conditions through constraint condition reduction, it can effectively reduce the complexity of the search algorithm and avoid the problems of the algorithm dependent on experience, so it is more versatile. However, due to the constraint condition reduction of the algorithm depends on the results of previous studies, it may lose precision when reducing the constraint conditions and reverse filter selection, how to avoid this problem will be the focus of future research.

REFERENCES

Banerjee, M., S. Mitra and H. Banka, 2007. Evolutionary-rough feature selection in gene expression data. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev., 37: 622-632.

Fan. L.L. and J. Wang, 2010. Clustering algorithms for mixed attributes based on rough set. J. Comput. Appl., 30: 3377-3379.

Jensen, R. and Q. Shen, 2007. Tolerance-based and fuzzy-rough feature selection. Proceedings of the IEEE International Conference on Fuzzy Systems, July 23-26, 2007, London, UK., pp: 877-882.

Momin, B.F., S. Mitra and R.D. Guota, 2006. Reduct generation and classification of gene expression data. Proceedings of the 1st International Conference on Hybrid Information Technology, November 9-11, 2006, New York, USA., pp: 699-708.

Parthalain, N.M. and Q. Shen, 2010. Exploring the boundary region of tolerance rough sets for feature selection. J. Comput. Eng. Appl., 46: 25-27.

Xu, S.M. and Q. Zhang, 2008. Optimization of DNA coding based on GA/PSO algorithm. Comput. Eng., 34: 218-220.

Yang, M. and P. Yang, 2008. A novel condensing tree structure for rough set feature selection. Neurocomputing, 71: 1092-1100.

Yao, Y. and Y. Zhao, 2009. Discernibility matrix simplification for constructing attribute reducts. Inform. Sci. Int. J., 179: 867-882.

Yin, Z., C.P. Ye and M. Wen, 2010. Research on DNA encoding design constraint by minimal free energy. Comput. Eng. Appl., 46: 25-27.

Zhang, B.H., H.S. Wang and L. Xu, 2006. Codes of DNA primary sequences and the similarity calculation. Chem. J. Chin. Univ., 27: 2277-2280.

Zhu, X.O., W.B. Liu and C. Sun, 2006. Research on the DNA words and algorithm. Acta Electron. Sin., 34: 1169-1174.