# SOLVING MULTIPLE SEQUENCE ALIGNMENT PROBLEM UTILIZING INTEGER LINEAR PROGRAMMING

PUDIAHWAI ANTON WIBOWO

0303010281

UNIVERSITY OF INDONESIA

FACULTY OF MATHEMATICS AND SCIENCE

DEPARTMENT OF MATHEMATICS

DEPOK

2008

# SOLVING MULTIPLE SEQUENCE ALIGNMENT PROBLEM

# UTILIZING INTEGER LINEAR PROGRAMMING

**This *skripsi* is submitted in partial fulfillment**
**of the requirements for the degree of**
***Sarjana Sains***

**By:**
**PUDIAHWAI ANTON WIBOWO**
**0303010281**

**DEPOK**
**2008**

Dedicated for my beloved parents,
Hwalianto, S.H and Diana Christianti,
upon their unconditional love.

“*Sikap sama dengan tujuan*.”
-My Father-

*SKRIPSI* : SOLVING MULTIPLE SEQUENCE ALIGNMENT PROBLEM UTILIZING INTEGER LINEAR PROGRAMMING.

NAME : PUDIAHWAI ANTON WIBOWO

NPM : 0303010281

THIS *SKRIPSI* HAS BEEN APPROVED

DEPOK, JULY 17th, 2008

| | |
|---|---|
| Dra. Denny R. Silaban, M.Kom | Dr. Kiki A. Sugeng, M.Si |
| ADVISOR I | ADVISOR II |

Date of completion of *sidang sarjana* examination: July 17th, 2008

Examiner I : Dra. Denny R. Silaban, M.Kom

Examiner II : Bevina D. Handari, Ph.D

Examiner III : Dra. Yahma Wisnani, M.Kom

# ABSTRACT

One of the dominant problems in computational molecular biology is multiple sequence alignment (MSA) of DNA. Many methods have been proposed to solve MSA problem such as dynamic programming and heuristic. A method has been proposed by Althaus *et al*. to solve MSA problem which is based on integer linear programming (ILP). The general ILP formulation of the MSA is derived from the graph representation of the MSA problem. Although we have the general ILP formulation of the MSA problem, constructing the ILP model of an MSA that can be solved directly using an ILP solver is not straightforward. We develop a MATLAB program that can generate and solve the ILP model of an MSA problem. The method that is used to solve the ILP model is branch-and-bound. The constructed program can generate the ILP model of any given MSA problem but can only solve an MSA problem of a small number of short DNA sequences. The result of the program is the aligned sequences of the MSA problem.

**Keywords**: MSA problem of DNA sequences, gapped extended alignment graph, gapped trace, ILP, branch-and-bound method.

xiv + 115 p.

Bibliography: 8 (1981 - 2006)

# ABSTRAK

Salah satu dari masalah-masalah dominan pada komputasi biologi molekuler adalah penyejajaran barisan berganda (*Multiple Sequence Alignment* - MSA) dari DNA. Banyak metode yang telah diajukan untuk menyelesaikan masalah MSA seperti pemrograman dinamik dan heuristik. Satu metode telah diajukan oleh Althaus *et al*. untuk menyelesaikan masalah MSA yang didasarkan pada pemrograman linear bilangan bulat (*Integer Linear Programming* - ILP). Formulasi ILP umum dari masalah MSA diturunkan dari representasi graf dari masalah MSA. Walaupun formulasi ILP umum dari masalah MSA diketahui, membentuk model ILP dari suatu masalah MSA yang dapat diselesaikan langsung menggunakan suatu *solver* ILP tidaklah mudah. Sebuah program yang dapat membangun dan menyelesaikan model ILP dari sebuah masalah MSA menggunakan MATLAB telah dibuat. Metode yang digunakan untuk menyelesaikan model ILP tersebut adalah branch-and-bound. Program yang telah dibuat dapat menghasilkan model ILP dari sembarang masalah MSA yang diberikan tetapi hanya dapat menyelesaikan masalah MSA dari sejumlah kecil barisan-barisan DNA yang pendek. Hasil dari program tersebut adalah penejajaran barisan-barisan DNA dari masalah MSA yang diberikan.

**Kata kunci**: masalah MSA barisan-barisan DNA, *gapped extended alignment graph*, *gapped trace*, ILP, *branch-and-bound method*.

xiv + 115 p.

Bibliografi: 8 (1981 - 2006)

Solving Multiple..., Pudiahwai Anton Wibowo, FMIPA UI, 2008

## CONTENTS

ACKNOWLEDGEMENT .......................................................................i

ABSTRACT ....................................................................................... iii

ABSTRAK ...........................................................................................v

CONTENTS ........................................................................................ vii

LIST OF FIGURE ...................................................................................ix

LIST OF TABLE ...................................................................................... xi

LIST OF APPENDIX ................................................................................ xii

CHAPTER I INTRODUCTION .......................................................1

1.1 BACKGROUND .............................................................. 1

1.2 PROBLEM STATEMENT ................................................ 3

1.3 THE PURPOSE OF THE *SKRIPSI* .................................. 3

1.4 SCOPE OF DISSCUSION ................................................4

1.5 THE ORGANIZATION OF THE *SKRIPSI* ........................ 4

CHAPTER II BASIC THEORIES ................................................................. 5

2.1 DEFINITIONS AND BASIC CONCEPTS OF MSA ...........6

2.2 SCORING SCHEME OF AN ALIGNMENT ...................... 9

2.3 DEFINITIONS AND BASIC CONCEPTS IN INTEGER LINEAR PROGRAMMING ............................................12

2.4 DEFINITIONS AND BASIC CONCEPTS IN GRAPH THEORY ......................................................................15

2.5 GRAPH REPRESENTATION OF THE MSA PROBLEM 27

2.6 BRANCH-AND-BOUND METHOD ................................ 29

CHAPTER III ILP MODEL OF THE MSA PROBLEM ................................ 35

3.1 THE GENERAL ILP FORMULATION OF THE MSA PROBLEM.................................................................. 36

3.2 STEPS OF GENERATING AN ILP MODEL OF AN MSA PROBLEM ................................................................. 41

3.2.1 DEFINING THE ALIGNMENT VARIABLES ......... 41

3.2.2 DEFINING THE GAP VARIABLES ...................... 44

3.2.3 DERIVING THE OBJECTIVE FUNCTION ........... 47

3.2.4 DERIVING THE FIRST TYPE CONSTRAINTS ... 50

3.2.5 DERIVING THE SECOND TYPE CONSTRAINTS 53

3.2.6 DERIVING THE THRID TYPE CONSTRAINTS .. 60

3.2.7 DERIVING THE FOURTH TYPE CONSTRAINTS 63

CHAPTER IV PROGRAM IMPLEMENTATION AND SIMULATION .......... 67

4.1 PROGRAM IMPLEMENTATION ................................... 67

4.2 PROGRAM SIMULATION ............................................ 77

CHAPTER V CONCLUSIONS .................................................................. 87

REFERENCES ........................................................................................... 89

APPENDIX ................................................................................................ 91

# LIST OF FIGURE

Fig. 2.1. Alignment set $\overline{S}$ .......................................................................... 7

Fig. 2.2. Possible alignments of Example 2.1 ............................................... 9

Fig. 2.3. Graph examples ........................................................................... 14

Fig. 2.4. Alignment graph of Example 2.1 .................................................. 15

Fig. 2.5. Extended alignment graph of Example 2.1 ................................... 16

Fig. 2.6. Gapped extended alignment graph of Example 2.1 ....................... 17

Fig. 2.7. Gapped trace of Figure 2.7 that corresponds to alignment (a) of Figure 2.2 .......................................................................... 20

Fig. 2.8. Examples of a mixed cycle and a critical mixed cycle .................... 23

Fig. 2.9. Contradictory ordering of mixed cycle Figure 2.8 (c) and (d) ......... 24

Fig. 2.10. Example of overlapping gap arcs of set $A_g^{3,1}$ ............................... 25

Fig. 2.11. Example of transitivity ................................................................. 26

Fig. 2.12. Branch-and-bound tree of Example 2.2 ....................................... 33

Fig. 3.1. Subgraph of Figure 2.6 that contains alignment edges of set $E^{1,3}$ . 42

Fig. 3.2. Subgraph of Figure 2.6 that contains all the possible gap arcs from sequence 1 .......................................................................... 44

Fig. 3.3. Example of incompatibility between alignment edge and gap arc .. 51

Fig. 3.4. Examples of critical mixed cycles .................................................. 54

Fig. 3.5. Subgraph of Figure 2.6 which contains all alignment edges that are incident to node $v_1^1$ and $v_2^1$ .......................................................... 56

Fig. 3.6. Figure 3.4 (d) with addition of a realized alignment edge by transitivity ............................................................................ 58

Fig. 3.7. Examples of incompatibility between gap arcs .............................. 62

Fig. 3.8. A cycle representing transitivity ..................................................... 62

Fig. 4.1. Initial MATLAB command window display of main.m .................... 68

Fig. 4.2. Display of an external input .......................................................... 68

Fig. 4.3. Example 2.1.txt ............................................................................. 69

Fig. 4.4. Display of text file path input ......................................................... 69

Fig. 4.5. Display of the result of MSA problem given in Example 2.1.txt ..... 71

Fig. 4.6. Display of the predefined input ..................................................... 72

Fig. 4.7. Display of the result of MSA problem 2 ......................................... 73

Fig. 4.8. Display of the result of MSA problem 3 ......................................... 74

Fig. 4.9. Display of the inputted DNA sequences ATG, ACG, TCG ............ 75

Fig. 4.10. Display of the result of Example 4.1 ............................................. 76

Fig. 4.11. Display of the result of Example 4.2 ............................................ 78

Fig. 4.12. Display of the result of Example 4.3 ............................................ 79

Fig. 4.13. Display of the result of Example 4.4 ............................................ 81

Fig. 4.14. Display of the result of Example 4.5 ............................................ 83

Fig. 4.15. Display of the result of Example 4.6 ............................................ 84

# LIST OF TABLE

Table 3.1. Total number of variables vs. increase in length of the sequences ........................................................................... 46

Table 3.2. Total number of variables vs. increase in number of the sequences ........................................................................... 46

Table 3.3. Total number of the first type constraints vs. increase in length of the sequences ..................................................................... 52

Table 3.4. Total number of the first type constraints vs. increase in number of the sequences ..................................................................... 52

Table 3.5. Total number of the second type constraints vs. increase in length of the sequences ................................................................. 60

Table 3.6. Total number of the second type constraints vs. increase in number of the sequences ..................................................... 60

Table 3.7. Total number of the third type constraints vs. increase in length of the sequences ..................................................................... 63

Table 3.8. Total number of the third type constraints vs. increase in number of the sequences ................................................................. 63

Table 3.9. Total number of the fourth type constraints vs. increase in length of the sequences ..................................................................... 66

Table 3.10. Total number of the fourth type constraints vs. increase in number of the sequences ..................................................... 66

Table 4.1. Running time of the ILP solver vs. increase in the length of the sequences ........................................................................ 80

Table 4.2. Running time of the ILP solver vs. increase in the dissimilarities of the sequences ................................................................... 85

# LIST OF APPENDIX

Appendix 1: The complete list of alignment variables for Example 2.1 ........ 91

Appendix 2: The complete list of gap variables for Example 2.1 .................. 91

Appendix 3 The complete objective function for Example 2.1....................... 92

Appendix 4: The complete first type constraints for Example 2.1 ................ 92

Appendix 5: The complete second type constraints for Example 2.1 ........... 94

Appendix 6: The complete third type constraints for Example 2.1 ............. 101

Appendix 7: The complete fourth type constraints for Example 2.1 ........... 102

Appendix 8: ILP solution of Example 2.1 .................................................. 105

Appendix 9: MSA – ILP source code ........................................................ 106