

ระบบออกแบบดีเจนนอเรทไฟรเมอร์เพื่อการศึกษาความหลากหลายทางชีวภาพของยีน  
โดยใช้การจับคู่แบบรูปพลวัต

นายวิวิศ ตีร์รัตนจากรู

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย  
ปีการศึกษา 2554  
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)  
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ที่ส่งผ่านทางบัณฑิตวิทยาลัย

The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)  
are the thesis authors' files submitted through the Graduate School.

Degenerate Primer Designing System for Gene Biodiversity Study  
using Dynamic Pattern Matching

Mr. Weeris Treeratanajaru

A Thesis Submitted in Partial Fulfillment of the Requirements  
for the Degree of Master of Science Program in Computer Science and Information  
Department of Mathematics and Computer Science  
Faculty of Science  
Chulalongkorn University  
Academic Year 2011  
Copyright of Chulalongkorn University

Thesis Title                    DEGENERATE PRIMER DESIGNING SYSTEM FOR GENE  
   BIODIVERSITY STUDY USING DYNAMIC PATTERN  
   MATCHING

By                                    MR. WEERIS TREERATANAJARU

Field of Study                    Computer Science and Information

Thesis Advisor                    Assistant Professor Rajalida Lipikorn, Ph.D.

Thesis Co-advisor                Supawin Watcharamul, Ph.D.

---

Accepted by the Faculty of Science, Chulalongkorn University in Partial  
Fulfillment of the Requirements for the Master's Degree

..... Dean of the Faculty of Science  
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

..... Chairman  
(Assistant Professor Nagul Cooharajanone, Ph.D.)

..... Thesis Advisor  
(Assistant Professor Rajalida Lipikorn, Ph.D.)

..... Thesis Co-advisor  
(Supawin Watcharamul, Ph.D.)

..... External Examiner  
(Associate Professor Rosarin Wongvilairat, Ph.D.)

วีรศ ตรีรัตนจารุ : ระบบออกแบบดีเจเนอเรทไพรเมอร์เพื่อการศึกษาความหลากหลายทางชีวภาพของยีน โดยใช้การจับคู่แบบรูปพลวัต. (DEGENERATE PRIMER DESIGNING SYSTEM FOR GENE BIODIVERSITY STUDY USING DYNAMIC PATTERN MATCHING) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.รัชติดา ลิปิกรณ์,อ. ที่ปรึกษาวิทยานิพนธ์ร่วม : อ.ดร.ศุภวิน วัชรมูล, 82 หน้า.

ในทางอณูชีววิทยา การออกแบบดีเจเนอเรทไพรเมอร์เพื่อการศึกษาความหลากหลายทางชีวภาพเป็นขั้นตอนที่ยาก ไพรเมอร์ต้องมีความเฉพาะเจาะจงกับยีนที่ต้องการศึกษา การเลือกไพรเมอร์ที่เหมาะสมนั้นเป็นเรื่องยาก และบางครั้งยังนำไปสู่การจับคู่แบบผิดตำแหน่ง ระบบออกแบบไพรเมอร์จำนวนมากถูกพัฒนาขึ้นเพื่อแก้ไขปัญหาในการศึกษาครั้งนี้ได้พัฒนาและประยุกต์ใช้การจับคู่แบบรูปพลวัตเพื่อค้นหายูนิเวอร์แซลไพรเมอร์ของแบคทีเรียและไพรเมอร์ที่เฉพาะเจาะจงต่อกลุ่มยีนโกลโคไซด์ ไฮโดรเลส กลุ่มที่ 45 โดยการจัดเรียงลำดับสารพันธุกรรมและส่งเข้าสู่ระบบ ซึ่งแบ่งเป็น 3 ขั้นตอน ได้แก่ การจัดเรียงข้อมูลใหม่ การออกแบบไพรเมอร์ และการคัดกรองคุณสมบัติ ขั้นแรกคือการคำนวณลำดับดีเจเนอเรทและลำดับคอนเซนซัสโดยใช้แบบจำลองทางสถิติ จากนั้นใช้ผลที่ได้ร่วมกับค่าพลังงานอิสระของกิบส์เพื่อออกแบบและเลือกลำดับที่เหมาะสมที่สุด เพื่อใช้เป็นลำดับของไพรเมอร์ นอกจากนี้ผู้วิจัยยังสามารถปรับแต่งคุณสมบัติต่างๆของไพรเมอร์ได้จากผลการทดลองพบว่า ดีเจเนอเรทไพรเมอร์ที่ออกแบบจากระบบนี้สามารถใช้งานได้จริง

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต.....  
 สาขาวิชา วิทยาการคอมพิวเตอร์สารสนเทศ..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์.....  
 ปีการศึกษา 2554..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

# # 5373605423 : MAJOR COMPUTER SCIENCE AND INFORMATION

KEYWORDS : DEGENERATER PRIMER DESIGN / GENE BIODIVERSITY / DYNAMIC PATTERN MATCHING

WEERIS TREERATANAJARU : DEGENERATE PRIMER DESIGNING SYSTEM FOR GENE BIODIVERSITY STUDY USING DYNAMIC PATTERN MATCHING.

ADVISOR : ASST. PROF. RAJARIDA LIPIKORN, Ph.D., CO-ADVISOR :

SUPAWIN WATCHARAMUL, Ph.D., 82 pp.

Degenerate primer design for studying biodiversity is a difficult step in molecular biology. The primers must be specific to the gene. Choosing the right primers are somewhat difficult and sometimes led to mismatching. Several primer design systems have been developed to overcome this problem. In this study, the Dynamic Pattern Matching technique has been developed and applied to find bacterial universal primers and specific primers for Glycoside Hydrolase Family 45 genes. Aligned sequences from test sets are entered into the system which consists of three steps: data reformation, primer design, and property filtering. First, degenerate and consensus sequences are calculated using statistical models. The results are combined with Gibbs Free Energy to design and select the most appropriate sequences as a series of primer sets. Moreover, users can also adjust their own criteria for each primer set. The results indicate that the degenerate primers designed by our proposed system are proved to be positive.

Department : Mathematics and Computer Science Student's Signature .....

Field of Study : Computer Science and Information Advisor's Signature .....

Academic Year : 2011 Co-advisor's Signature .....

## ACKNOWLEDGEMENTS

I would like to acknowledge my advisor and co-advisor, Assistant Professor Dr. Rajalida Lipikorn, at Department of Mathematics and Computer Science, and Dr. Supawin Watcharamul, at Department of Environmental Science, Faculty of Science, Chulalongkorn University, for all their great support helping me to improve my researching skill, and giving me the great opportunity in publishing proceedings.

I would like to thank Assistant Professor Dr. Roongkan Nuisin at Department of Environmental Science, Faculty of Science, Chulalongkorn University and Miss Phasinee Khwanmuang for information support in thermodynamic part.

In addition, the research will not be complete without the main support from the Centre of Excellent in Mathematics, CHE, Sri Ayutthaya Rd., Bangkok, 10400, Thailand. The Center of Excellent in Mathematics brings the supports via "Research Assistants scholarship." I would like to appreciate that great support for necessary tuition fees, and other education costs.

I would like to thank Miss Jakwida Choowongsirikul for friendly and helpful support in dissertation work.

Finally I would like to thank my father, mother and friends for everything they suggested and supported me.

## CONTENTS

	PAGE
ABSTRACT (THAI).....	iv
ABSTRACT (ENGLISH).....	v
ACKNOWLEDGEMENTS.....	vi
CONTENTS.....	vii
LIST OF TABLES.....	x
LIST OF FIGURES.....	xii
CHAPTER	
<b>I INTRODUCTION.....</b>	<b>1</b>
1.1 Objectives.....	3
1.2 Scope.....	3
1.3 Research methodology.....	4
1.4 Expected outcomes.....	5
<b>II LITERATURE REVIEW.....</b>	<b>6</b>
2.1 Thermodynamic approach for primer design.....	6
2.1.1 Melting temperature.....	6
2.1.2 Primer length.....	6
2.1.3 GC content.....	7
2.1.4 Single repeat.....	7
2.1.5 Di-repeat.....	7
2.1.6 Secondary structures.....	7
2.2 Primer designing system.....	8
2.2.1 CODEHOP and iCODEHOP.....	8
2.2.2 GeneFisher and GeneFisher-P.....	10
2.2.3 PaBaLis.....	11

CHAPTER	PAGE
2.2.4 HYDEN.....	11
2.2.5 UniPrime and UniPrime2.....	12
2.2.6 Greene SCPrimer .....	13
2.2.7 PerlPrimer.....	13
2.2.8 MAD-DPD.....	13
2.2.9 Primique.....	14
2.2.10 Primaclade.....	14
2.2.11 Amplicon.....	14
2.2.12 PAMPS.....	15
2.2.13 PCR-RFLP primer design using genetic algorithm.....	15
2.2.14 Optimus Primer.....	15
2.2.15 GC-rich primer design strategy.....	16
2.2.16 Primer3.....	17
<b>III PROPOSED METHOD .....</b>	<b>22</b>
3.1 Data reformation.....	23
3.2 Primer design.....	27
3.3 Property filtering.....	28
<b>IV EXPERIMENT RESULTS AND DISCUSSION.....</b>	<b>29</b>
4.1 PCR primer evaluation.....	29
4.1.1 Bacteria universal primer designing.....	31
4.1.1.1 The first experiment.....	31
4.1.1.2 The second experiment.....	35
4.1.1.3 The third experiment.....	39
4.1.1.4 The fourth experiment.....	44
4.1.2 GH45 specific primer designing.....	46
4.2 Computational complexity evaluation.....	48



CHAPTER	PAGE
V CONCLUSION.....	49
REFERENCES.....	51
APPENDICES.....	55
Appendix A Sequences of 16s rRNA gene of 44 bacteria species.....	56
Appendix B Amino acid sequences of glycoside hydrolase family 45....	74
Appendix C Installation of CUPrimer program.....	75
Appendix D CUPrimer program manual guide.....	79
BIOGRAPHY.....	82

## LIST OF TABLES

TABLE	PAGE
1.1 IUPAC nomenclature of mixed bases.....	2
1.2 Task schedule.....	4
2.1 Comparison of present primer designing programs (Biological view).....	18
2.2 Comparison of present primer designing programs (Computational view)..	20
3.1 Example sequences.....	23
3.2 Nucleotide Matrices (NM).....	23
3.3 Anti-Nucleotide Matrices (ANM).....	24
3.4 Degenerate Matrices (DM).....	24
3.5 Anti-Degenerate Matrices (ADM).....	25
3.6 Consensus Matrices (CM).....	26
3.7 Anti-Consensus Matrices (ACM).....	26
3.8 Sequence generated from ANM.....	27
4.1 26 universal primers for result comparison from proposed system.....	30
4.2 Degenerate primer for GH 45.....	27
4.3 16s rRNA genes of 44 organisms for system testing (first experiment).....	27
4.4 The first experiment results.....	31
4.5 16s rRNA genes of 41 organisms for system testing (second experiment)...	35
4.6 The second experiment results.....	38
4.7 Universal primers designed in the second experiment.....	38
4.8 $\Delta G$ and $3'\Delta G$ of universal primer output from second experiment.....	39
4.9 16s rRNA genes of 37 organisms for system testing (third experiment).....	40
4.10 The third experiment results.....	42
4.11 Universal primers designed in the third experiment.....	42
4.12 16s rRNA genes of 10 organisms for system testing (fourth experiment)...	44
4.13 The fourth experiment results.....	45
4.14 Universal primers designed in the fourth experiment.....	45

TABLE	PAGE
4.15 Amino acid sequence of 5 organisms for system testing.....	46
4.16 Specific primer designing experiment result.....	47

## LIST OF FIGURES

FIGURE	PAGE
3.1 Process of the proposed degenerate primer designing system.....	22
3.2 Dynamic Pattern Matching (DPM) work flow.....	28
4.1 Example of a sequence in FASTA format.....	29
4.2 The first experiment criteria setting.....	34
4.3 The first experiment aligned sequence.....	35
4.4 The second experiment criteria setting.....	38
4.5 The third experiment criteria setting.....	42
4.6 The third experiment aligned sequence.....	43
4.7 The fourth experiment criteria setting.....	45
4.8 Specific primer designing experiment criteria setting.....	47

# CHAPTER I

## INTRODUCTION

In molecular biology, many new techniques have been used to study the diversity of target genes in environment. Most frequently used technique is Polymerase Chain Reaction (PCR) which can amplify a specific region of DNA such that enough copies of that region are available to be adequately tested or sequenced [1]. In order to use PCR, one must know the exact sequences which lie on either side of the DNA region of interest. These sequences are used to design a set of nucleotide sequences called *primers*, complement to each strand of the DNA and lying on opposite side of the target region. The effectiveness of the technique depends on choosing the primer sets. Primer design must be used to fit specific genes.

When some of the sequence positions contain several possible bases [2] at the exact position, this type of primer sequence is referred as *degenerate*. For example, in the primer GA{A,G}T{C,G,T}C, the third position is either A or G and the fifth is either C, G, or T. The IUPAC illustration will be GARTBC as shown in **Table 1.1**. The degeneracy of a primer is the number of unique sequence combinations it contains. For example, the degeneracy of the above primer is six. Degenerate primers can accommodate all possible codons of all amino acid residuals [3].

To design sets of primers that cover diversity of genes and fortunately a new gene, one should obviously use primers that have high degeneracy. On the other hand, in order to reduce the probability of amplifying unrelated genes, the degeneracy must be bounded. This contradictory nature of the degenerate primer design problem has led to definition of several variants of this problem, all of which are NP-complete [4].

Refer to related works about degenerate primer design; Rose *et al.* (1998) proposed COnsensus-DEgenerate Hybrid Oligonucleotide Primer (CODEHOP) strategy, which consists of a short 3' degenerate region and a longer 5' consensus region. CODEHOP method has been proved positive by many researches [5]. This technique is

fixed region technique that suitable for low diversity gene study. In the other hand, for high diversity gene study, Linhart and Shamir (2002) implemented technique for design highly-degenerate primer call HYDEN, which uses approximation algorithms for solve the various simplified problems [4]. This technique requires complicated processing. The other technique used for degenerate primer designing is applying flexible region length primer to identify and clone a large group of gene. For this technique, Pan *et al.* (2007) reported PaBaLiS strategy for design hybrid partially degenerate primers which consisted of three flexible primer regions [6].

**Table 1.1:** IUPAC nomenclature of mixed bases

IUPAC nucleotide	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

In this thesis, appropriate features will be proposed along with thermodynamic approach for hybrid degenerate primer design that contain the flexible region of degenerate and consensus sequence regions for yielding high performance in term of accuracy and mismatch acceptance. To evaluate our features and system for hybrid degenerate primer design, the proposed system result was compared with existing universal primer for bacterial DNA and specific degenerate primer of cellulose genes. For the proposed techniques, the input is alignment sequence that contains part of target sequence. Dynamic Pattern Matching (DPM) algorithm is proposed as a new technique to handle problems of having non-specificity in choosing genes and better running time.

This thesis is organized as follows. The second chapter explains the related works, thermodynamic approach and degenerate primer designing techniques which are applied. The third chapter shows the proposed method. The fourth chapter is about experimental results and discussion. The last chapter is conclusion. Some parts of this work contain the details which were already proposed in the proceeding of the 7th International Symposium on Health Informatics and Bioinformatics 2012 (HIBIT2012).

### 1.1 Objectives

The main objective of this study is developing primer designing system for specific gene biodiversity study.

### 1.2 Scope

In this study, the primer designing system is constrained as follows:

1. Input sequences are any DNA or amino acid sequence.
2. Alignment sequence is allowed in .aln format.





#### 1.4 Expected outcomes

1. Computer based system for design degenerate primer that can be used for specific gene biodiversity.
2. New technique for design degenerate primer based on thermodynamic parameter.

## CHAPTER II

### LITERATURE REVIEW

Primer designing system is the challenge issue for bioinformatics field. 10-20 primer criteria must be optimize. Many researcher study in this field and develop the new techniques for optimize those criteria. Thermodynamic is the core of most programs however there are another 2-3 theory can use to select the best primer sets for PCR.

#### 2.1 Thermodynamic approach for primer design

Many techniques were used for design primer sets within some criteria. The basic one is string matching algorithm that apply the basic computer problem to use for solve biological problem. String matching algorithm can show the good running time but it cannot handle the complexity of genetic sequence such as amino acid sequence. For more accurate, thermodynamic concept was used as core of most systems.

The important design considerations described below are a key to specific amplification with high yield. The preferred values indicated are built into all our products by default.

2.1.1. Melting Temperature ( $T_m$ ) by definition is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates the duplex stability. Primers with melting temperatures in the range of 52-58 °C generally produce the best results. Primers with melting temperatures above 65°C have a tendency for secondary annealing. The GC content of the sequence gives a fair indication of the primer  $T_m$ .

2.1.2. Primer length is generally accepted that the optimal length of PCR primers is 18-22 base pairs. This length is long enough for adequate specificity, and short enough for primers to bind easily to the template at the annealing temperature.

2.1.3. The GC content is the number of G's and C's in the primer as a percentage of the total bases of primer should be 40-60%.

2.1.4. Single repeat is an occurrence that primer has long runs of a single base should generally be avoided as they can generate the **secondary** structure. For example, AGCGGGGGATGGGG has single repeat of base 'G' of value five and four. A maximum number of single repeat accepted is 4 base pairs.

2.1.5. Di-repeat is an occurrence that primer has long runs of a pair base should generally be avoided as they can generate the secondary structure. For example: ATATATAT. A maximum number of di-nucleotide repeats acceptable in an oligo is four di-nucleotides.

2.1.6. Presence of the primer secondary structures produced by intermolecular or intramolecular interactions can lead to poor or no yield of the product. They adversely affect primer template annealing and thus the amplification. They greatly reduce the availability of primers to the reaction.

2.1.6.1. Hairpins: It is formed by intramolecular interaction within the primer and should be avoided. Optimally a 3' end hairpin with a  $\Delta G$  of -2 kcal/mol and an internal hairpin with a  $\Delta G$  of -3 kcal/mol is tolerated generally.

2.1.6.2. Self-Dimer: A primer self-dimer is formed by intermolecular interactions between the two (same sense) primers, where the primer is homologous to itself. Generally a large amount of primers are used in PCR compared to the amount of target gene. When primers form intermolecular dimers much more readily than hybridizing to target DNA, they reduce the product yield. Optimally a 3' end self-dimer with a  $\Delta G$  of -5 kcal/mol and an internal self-dimer with a  $\Delta G$  of -6 kcal/mol is tolerated.

2.1.6.3. Cross-Dimer: Primer cross dimers are formed by intermolecular interaction between sense and antisense primers, where they are

homologous. Optimally a 3' end cross-dimer with a  $\Delta G$  of -5 kcal/mol and an internal cross-dimer with a  $\Delta G$  of -6 kcal/mol is tolerated generally.

2.1.7. Primer Pair  $T_m$  Mismatch Calculation, the two primers of a primer pair should have closely matched melting temperatures for maximizing PCR product yield. The difference of 5°C or more can lead no amplification.

## 2.2 Primer designing system

Many programs were developed for some specific propose of PCR technique for example high GC content region amplification, high degeneracy genes amplification and primer designing large number of sequences. In this topic explains advantages and disadvantages of each technique and highlight of each program.

### 2.2.1 CODEHOP and iCODEHOP

Rose *et al.* (1998) presented a new primer design technique for PCR amplification of unknown targets that are related to multiply-aligned protein sequences. Each primer consists of a short degenerate core region and a longer consensus clamp region. Only 3–4 highly conserved amino acid sequences are necessary for design of the core, which is stabilized by the clamp during annealing to template molecules. During later rounds of amplification, the non-degenerate clamp permits stable annealing to product molecules. This COnsensus-DEgenerate Hybrid Oligonucleotide Primer (CODEHOP) strategy has been implemented as a computer program that is accessible over the World Wide Web (<http://blocks.fhcrc.org/codehop.html>) and is directly linked from the BlockMaker multiple sequence alignment sites for hybrid primer prediction beginning with a set of related protein sequences.

Rose *et al.* (2003) prove that CODEHOPs can be used in PCR amplification to isolate distantly related sequences encoding the conserved amino acid sequence. The primer design software and the CODEHOP PCR strategy have been utilized for the identification and characterization of new gene orthologs and paralogs in different plant,

animal and bacterial species. In addition, this approach has been successful in identifying new pathogen species.

Staheli *et al.* (2009) reported that CODEHOP have proven to be a powerful tool for the identification of novel genes. The CODEHOPs approach has been used to identify novel pathogens by targeting amino acid motifs conserved in specific pathogen families. They initiated a program utilizing the CODEHOPs approach to develop PCR-based assays targeting and further improved a computer program and website to facilitate the design of CODEHOPs PCR primers. They detail the method for the development of pathogen-specific CODEHOPs PCR assays using the papillomavirus family as a target. Papillomaviruses constitute a diverse virus family infecting a wide variety of mammalian species, including human and non-human primates. We demonstrate that our pan-papillomavirus CODEHOPs assay is broadly reactive with all major branches of the virus family and show its utility in identifying a novel non-human primate papillomavirus in cynomolgus macaques.

Staheli *et al.* (2011) studied about genes that are supposed to be distantly related to a known gene sequence, such as homologous genes in different species, paralogs in the same genome, or novel pathogens in diverse hosts, often turns into the proverbial search for the needle in the haystack. PCR-based methods commonly used to address this issue involve the use of either consensus primers or degenerate primers, both of which have significant shortcomings regarding sensitivity and specificity. They have developed a novel primer design approach that diminishes these shortcomings and instead takes advantage of the strengths of both consensus and degenerate primer designs, by combining the two concepts into a Consensus–Degenerate Hybrid Oligonucleotide Primer (CODEHOP) approach. During the initial PCR amplification cycles, the degenerate core is responsible for specific binding to sequences encoding the conserved amino acid motif. The longer consensus clamp region serves to stabilize the primer and allows the participation of all primers in the pool in the efficient amplification of products during later PCR cycles. They have developed an interactive

web site and algorithm (iCODEHOP) for designing CODEHOP PCR primers from multiply aligned protein sequences, which is freely available online.

### 2.2.2 GeneFisher and GeneFisher-P

Giegerich and F. Meyer (1996) proposed GeneFisher program for study about a family of genes from closely related organisms is known, there is a certain chance to extract the corresponding gene from the genome of another related organism. This can be done by polymerase chain reaction, provided that a pair of suitable primers can be designed. In contrast to primer design for a single, known target sequence, systematic primer design for an unknown target given a group of homologues can by no means be done manually. GeneFisher is a software tool which automates this task, and takes special care to make the impact of the manifold design parameters transparent to the user.

Lamprecht *et al.* (2008) describe the popular tool GeneFisher and explain its recent restructuring using workflow techniques. They apply a service-oriented approach to model and implement GeneFisher-P, a process-based version of the GeneFisher web application, as a part of the Bio-jETI platform for service modelling and execution. They show how to introduce a flexible process layer to meet the growing demand for improved user-friendliness and flexibility. Within Bio-jETI, model the process using the jABC framework, a mature model driven service-oriented process definition platform. They encapsulate remote legacy tools and integrate web services using jETI, an extension of the jABC for seamless integration of remote resources as basic services, ready to be used in the process. Some of the basic services used by GeneFisher are in fact already provided as individual web services at BiBiServ and can be directly accessed. Others are legacy programs, and are made available to Bio-jETI via the jETI technology. The full power of service-based process orientation is required when more bioinformatics tools, available as web services or via jETI, lead to easy extensions or variations of the basic process. This concerns for instance variations of data retrieval or alignment tools as provided by the EBI. The resulting service- and process-oriented

GeneFisher-P demonstrates how basic services from heterogeneous sources can be easily orchestrated in the Bio-jETI platform and lead to a flexible family of specialized processes tailored to specific tasks.

### 2.2.3 PaBaLis

Pan *et al.* (2007). report a novel and successful selection strategy for the design of hybrid partially degenerate primers for use with RT-PCR and RACE-PCR for the identification of unknown gene families. The technique (named PaBaLiS) has proven very effective as it allowed us to identify and clone a large group of mRNAs encoding neurotoxin-like polypeptide pools from the venom of *Agelena orientalis* species of spider. Their approach differs radically from the generally accepted CODEHOP principle first reported in 1998. Most importantly, our method has proven very efficient by performing better than an independently generated high throughput EST cloning programme. Their method yielded nearly 130 non-identical sequences from *Agelena orientalis*, whilst the EST cloning technique yielded only 48 non-identical sequences from 2100 clones obtained from the same *Agelena* material. In addition to the primer design approach reported here, which is almost universally applicable to any PCR cloning application, our results also indicate that venom of *Agelena orientalis* spider contains a much larger family of related toxin-like sequences than previously thought. With upwards of 100,000 species of spider thought to exist, and a propensity for producing diverse peptide pools, many more peptides of pharmacological importance await discovery. They envisage that some of these peptides and their recombinant derivatives will provide a new range of tools for neuroscience research and could also facilitate the development of a new generation of analgesic drugs and insecticides.

### 2.2.4 HYDEN

Linhart and Shamir (2002, 2005) study the problem of designing a pair of primers with prescribed degeneracy that match a maximum number of given input sequences. Such problems occur when studying a family of genes that is known only in

part, or is known in a related species. We prove that various simplified versions of the problem are hard, show the polynomiality of some restricted cases, and develop approximation algorithms for one variant. Based on these algorithms, They implemented a program called HYDEN for designing highly-degenerate primers for a set of genomic sequences. They report on the success of the program in an experimental scheme for identifying all human olfactory receptor (OR) genes. In that project, HYDEN was used to design primers with degeneracies up to  $10^{10}$  that family, tripling the number of OR genes known at the time.

### 2.2.5 UniPrime and UniPrime2

Bekaert and Teeling (2008) report UniPrime that is an open-source software (<http://uniprime.batlab.eu>), which automatically designs large sets of universal primers by simply inputting a gene ID reference. UniPrime automatically retrieves and aligns homologous sequences from GenBank, identifies regions of conservation within the alignment and generates suitable primers that can amplify variable genomic regions. UniPrime differs from previous automatic primer design programs in that all steps of primer design are automated, saved and are phylogenetically limited. They have experimentally verified the efficiency and success of this program by amplifying and sequencing four diverse genes (AOF2, EFEMP1, LRP6 and OAZ1) across multiple Orders of mammals. UniPrime is an experimentally validated, fully automated program that generates successful cross-species primers that take into account the biological aspects of the PCR.

Boutros *et al.* (2009) report UniPrime2 web server which is a publicly available online resource which automatically designs large sets of universal primers when given a gene reference ID or Fasta sequence input by a user. UniPrime2 works by automatically retrieving and aligning homologous sequences from GenBank, identifying regions of conservation within the alignment, and generating suitable primers that can be used to amplify variable genomic regions. In essence, UniPrime2 is a suite of publicly available software packages (Blastn, T-Coffee, GramAlign, Primer3), which



reduces the laborious process of primer design, by integrating these programs into a single software pipeline. Hence, UniPrime2 differs from previous primer design web services in that all steps are automated, linked, saved and phylogenetically delimited, only requiring a single user-defined gene reference ID or input sequence. They provide an overview of the web service and wet-laboratory validation of the primers generated.

#### 2.2.6 Greene SCPPrimer

Jabado *et al.* (2006) presented a method for designing such primers based on tree building followed by application of a set covering algorithm, and demonstrate its utility in compiling Multiplex PCR primer panels for detection and differentiation of viral pathogens

#### 2.2.7 PerlPrimer

Marshall (2004) proposed PerlPrimer that is a cross-platform graphical user interface application for the design of primers for standard, bisulphite and real-time PCR, and sequencing. The program incorporates highly accurate melting-temperature and primer-dimer prediction algorithms with powerful tools such as sequence retrieval from Ensembl and the ability to BLAST search primer pairs. It aims to automate and simplify the process of primer design.

#### 2.2.8 MAD-DPD

Najafabadi *et al.* (2008) introduced minimum accumulative degeneracy, a variant of the degenerate primer design problem, which is particularly useful when a large number of sequences are to be covered by a set of restricted number of primers. A primer set, which is designed on a minimum accumulative degeneracy basis, especially helps to reduce nonspecific PCR amplification of undesired DNA fragments, as fewer primer species are present in PCR. A Boltzmann machine is designed to solve the minimum accumulative degeneracy degenerate primer design problem, called the MAD-DPD Boltzmann machine. This algorithm shows great flexibility, as it can be

determined either to solve the problem with strict fidelity to covering all input sequences or to exclude some input sequences if it results in less degenerate primers. This Boltzmann machine is successfully implemented in designing a new set of primers for amplification of antibody variable fragments from mouse spleen cells, which theoretically covers more diverse antibody sequences than currently available primers.

#### 2.2.9 Primique

Fredslund and Lange (2007) presented primique, a new graphical, user-friendly, fast, web-based tool which solves the problem: It designs specific primers for each sequence in an uploaded set. Further, a secondary set of sequences *not* to be amplified by any primer pair may be uploaded. Primers with high sequence similarity to non-target sequences are selected against. Lastly, the suggested primers may be checked against the National Center for Biotechnology Information databases for possible mis-priming. Results are presented in interactive tables, and various primer properties are listed and displayed graphically. Any close match alignments can be displayed. Given 30 sequences, the running time of primique is about 20 seconds.

#### 2.2.10 Primaclade

Gadberry *et al.* (2004) presented Primaclade which is a web-based application that accepts a multiple species nucleotide alignment file as input and identifies a set of polymerase chain reaction (PCR) primers that will bind across the alignment. Primaclade iteratively runs the Primer3 application for each alignment sequence and collates the results. Primaclade creates an HTML results page that recaps the original alignment, provides a consensus sequence and lists primers for each alignment area, with primers color-coded to reflect the level of degeneracy in the primer.

#### 2.2.11 Amplicon

Jarman (2004) presented Amplicon which is a program for designing PCR primers on aligned groups of DNA sequences. The most important application for

Amplicon is the design of 'group-specific' PCR primer sets that amplify a DNA region from a given taxonomic group but do not amplify orthologous regions from other taxonomic groups.

#### 2.2.12 PAMPS

Najafabadi *et al.* (2008) presented PAMPS, the method presented in this work, usually results in a 30% reduction in the number of degenerate primers required to cover all sequences, compared to the previous algorithms. In addition, PAMPS runs up to 3500 times faster. Due to small running time, using PAMPS allows designing degenerate primers for huge numbers of sequences. In addition, it results in fewer primers which reduces the synthesis costs and improves the amplification sensitivity.

#### 2.2.13 PCR-RFLP primer design using genetic algorithm

Yang *et al.* (2010) proposed the new technique for performing PCR-RFLP for SNP genotyping, a feasible primer pair, which must observe numerous constraints, and an available restriction enzyme for discriminating a target SNP, are required. Here, we propose a method which uses a genetic algorithm (GA) to search for optimal natural PCR-RFLP primers and employs the core of SNP-RFLPing to reliably mine available restriction enzymes. The *in silico* simulation of the proposed method in the SNPs of the SLC6A4 gene showed that it is able to stably to design natural PCR-RFLP primers which most fit the common primer constraints and provide available restriction enzymes.

#### 2.2.14 Optimus Primer

Brown *et al.* (2010) presented program Optimus Primer (OP) that automatically takes into account all these variables, and can generate primers with no need to provide genome coordinates. More importantly however, OP, unlike other primer design programs, uniquely utilizes Primer3 in an iterative manner that allows the user to progressively design up to four iterations of primer designs. Through a single interface, the user can specify up to four different design parameters with different stringencies,

thus increasing the probability that a functional PCR primer pair will be designed for all regions of interest in a single pass of the pipeline. To demonstrate the effectiveness of the program, we designed PCR primers against 77 genes located in loci associated with ulcerative colitis as part of a candidate gene re-sequencing experiment. We achieved an experimental success rate of 93% or 472 out of 508 amplicons spanning the exonic regions of the 77 genes. Moreover, by automatically passing amplicons that failed primer design through three additional iterations of design parameters, we achieved an additional 170 successful primer pairs or 34% more in a single pass of OP than by conventional methods. With only a gene list and PCR parameters, a user can produce hundreds of PCR primer designs for regions of interest with a high probability of success in a very short amount of time. Optimus Primer is an essential tool for researchers who want to pursue PCR-based enrichment strategies for next-generation re-sequencing applications.

#### 2.2.15 GC-rich primer design strategy

Li *et al.* (2011) established a primer design method for amplification of GC-rich DNA sequences. A group of 15 pairs of primers with higher  $T(m)$  ( $>79.7^{\circ}\text{C}$ ) and lower level  $\Delta T(m)$  ( $<1^{\circ}\text{C}$ ) were designed to amplify GC-rich sequences (66.0%-84.0%). The statistical analysis of primer parameters and GC content of PCR products was performed and compared with literatures. Other control experiments were conducted using shortened primers for GC-rich PCR amplifications in this study, and the statistical analysis of shortened primer parameters and GC content of PCR products was performed compared with primers not shortened. A group of 26 pairs of primers were designed to test the applicability of this primer designing strategy in amplifications of non-GC-rich sequences (35.2%-53.5%). All the DNA sequences in this study were successfully amplified. Statistical analyses show that the  $T(m)$  and  $\Delta T(m)$  were the main factors influencing amplifications. This primer designing strategy offered a perfect tool for amplification of GC-rich sequences. It proves that the secondary structures cannot

be formed at higher annealing temperature conditions ( $>65^{\circ}\text{C}$ ), and we can overcome this difficulty easily by designing primers and using higher annealing temperature.

#### 2.2.16 Primer3

Rozen and Skaletsky presented Primer3 is a computer program that suggests PCR primers for a variety of applications, for example to create STSs (sequence tagged sites) for radiation hybrid mapping, or to amplify sequences for single nucleotide polymorphism discovery. Primer3 can also select single primers for sequencing reactions and can design oligonucleotide hybridization probes. In selecting oligos for primers or hybridization probes, Primer3 can consider many factors. These include oligo melting temperature, length, GC content, 3' stability, estimated secondary structure, the likelihood of annealing to or amplifying undesirable sequences (for example interspersed repeats), the likelihood of primer-dimer formation between two copies of the same primer, and the accuracy of the source sequence. In the design of primer pairs Primer3 can consider product size and melting temperature, the likelihood of primer-dimer formation between the two primers in the pair, the difference between primer melting temperatures, and primer location relative to particular regions of interest or to be avoided.

From reviewing, there are many other researchers that do not mention about the mismatch of adjacent nucleotide that is the key of degenerate primer design. Hybrid primer concept such as PaBaLiS, CODEHOP and iCODEHOP are the most acceptable program. However, that program is the fixed region hybrid primer; mean that primers are separated into 2-3 regions for decrease degeneracy value. But in fact, degenerate primer algorithm such as Primer3 run without separation and also can found the good primer. In this work, we want to implement the new technique that can apply the hybrid concept without separation of primer that will be mentioned in the next chapter.

**Table 2.1:** Comparison of present primer designing programs (Biological view)

System name	Ref.	System properties						
		Thermodynamic criteria using	Degeneracy control	Multiplex PCR primer	Hybrid primer	Mismatch acceptance	Design from amino acid sequence	Special feature
CODEHOP	[5][7][8][9]	Yes	Yes	No	Yes	Yes	Yes	2 phases primer
GeneFisher	[10][11]	Yes	No	No	No	Yes	Yes	-
PaBaLiS	[12]	Yes	Yes	No	Yes	Yes	Yes	3 phases primer
HYDEN	[4][13]	No	No	Yes	No	Yes	No	-
UniPrime	[14][15]	Yes	Yes	No	No	Yes	No	-
Greene SCPrimer	[16]	No	Yes	Yes	No	Yes	No	-
PerlPrimer	[17]	Yes	No	No	No	Yes	No	-
MAD-DPD	[18]	No	Yes	Yes	No	Yes	No	-
Primique	[19]	Yes	Yes	No	No	Yes	Yes	-

Table 2.1 (cont.): Comparison of present primer designing programs (Biological view)

System name	Ref.	System properties						
		Thermodynamic criteria using	Degeneracy control	Multiplex PCR primer	Hybrid primer	Mismatch acceptance	Design from amino acid sequence	Special feature
Primaclade	[20]	N/A	Yes	No	No	No	No	Only conserved region found
Amplicon	[21]	N/A	Yes	No	No	N/A	No	Group-specific primers
PAMPS	[22]	No	Yes	Yes	No	Yes	No	-
PCR-RFLP primer design	[23]	No	Yes	No	No	Yes	No	SNP genotyping
Optimus Primer	[24]	No	No	No	No	Yes	No	Human exonic region specific
GC-rich primer	[25]	Yes	No	No	No	No	No	GC-rich
Primer3	[26]	Yes	No	No	No	Yes	No	

Table 2.2: Comparison of present primer designing programs (Computational view)

System name	Ref.	System properties					
		Algorithm	Running Time	Graphical	Freeware	Open source	Database connection
CODEHOP	[5][7][8][9]	Heuristic	N/A	No	Yes	No	Yes
GeneFisher	[10][11]	Heuristic	N/A	No	Yes	Yes	No
PaBaLiS	[12]	N/A	N/A	No	No	No	No
HYDEN	[4][13]	Heuristic	N/A	No	Yes	No	No
UniPrime	[14][15]	N/A	N/A	No	Yes	Yes	Yes
Greene SCPrimer	[16]	Greedy	$O(n^3)$	No	Yes	No	No
PerlPrimer	[17]	N/A	N/A	No	Yes	Yes	No
MAD-DPD	[18]	Heuristic	N/A	No	Yes	No	No
Primique	[19]	N/A	N/A	No	Yes	No	Yes
Primaclade	[20]	N/A	N/A	No	Yes	Yes	Yes
Amplicon	[21]	Heuristic	N/A	No	Yes	No	No
PAMPS	[22]	Heuristic	N/A	No	Yes	No	No
PCR-RFLP primer design	[23]	Genetic	$O(n^7)$	N/A	No	No	Yes
Optimus Primer	[24]	N/A	N/A	No	Yes	No	No



Table 2.2 (cont.): Comparison of present primer designing programs (Computational view)

System name	Ref.	System properties					
		Algorithm	Running Time	Graphical	Freeware	Open source	Database connection
GC-rich primer	[25]	N/A	N/A	N/A	N/A	N/A	N/A
Primer3	[26]	N/A	N/A	No	Yes	Yes	No

## CHAPTER III

### PROPOSED METHOD

In this thesis, the technique called Dynamic Pattern Matching (DPM) is proposed for designing degenerate primer. The proposed system consists of three steps: data reformation, primer design, and property filtering. Figure 3.1 shows the proposed system process.

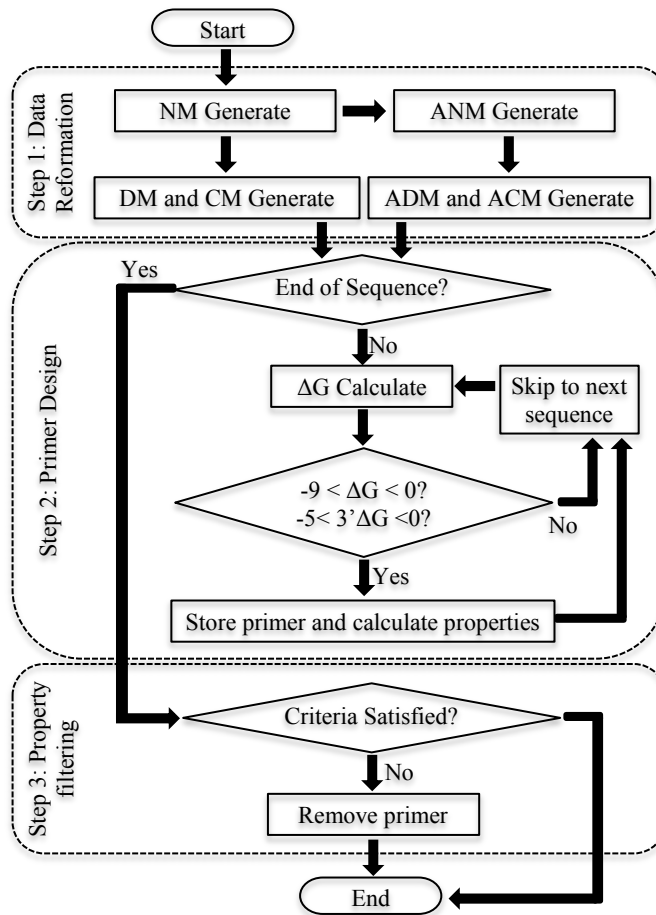


Figure 3.1. Process of the proposed degenerate primer designing system

### 3.1. Data reformation

In this step, an aligned sequence is transformed to probability of having each nucleotide base at each position by using a two dimensional arrays called a Nucleotide Matrices (NM), as expressed in Eq. (1) and Table 3.2

$$N_{n,i} = \frac{\sum_{j=1}^k m_{n,i,j}}{k} \quad (1)$$

where  $n$  represents nucleotide base (A,T,C,G),  $i$  indicates the position of nucleotide in an aligned sequence,  $N_{n,i}$  is the probability of having nucleotide  $n$  at position  $i$ ,  $k$  is the number of sequences and  $m_{n,i,j}$  is the number of nucleotide base  $n$  at position  $i$  in sequence  $j$ . If  $n$  is any other degenerate code or nucleotide then the probability of having nucleotide base  $n$  in that code is used instead. For example, a degenerate code B has a probability of being C, G or T equal to 0.33.

Table 3.1: Example sequences

Position	1	2	3	4	5
Seq. 1	A	T	C	G	A
Seq. 2	A	T	C	A	T
Seq. 3	A	T	C	A	C

Table 3.2: Nucleotide Matrices (NM) generated from sequences in Table 3.1

Position	1	2	3	4	5
A	1.00	0.00	0.00	0.67	0.33
T	0.00	1.00	0.00	0.00	0.33
C	0.00	0.00	1.00	0.00	0.33
G	0.00	0.00	0.00	0.33	0.00

The system then uses NM to generate an antisense strand DNA and transform it to an Anti-Nucleotide Matrices (ANM) by taking complement of  $N_{n,i}$  of each base pair and reversing the position,  $i$  as expressed in Eq. (2) and Table 3.3

$$A_{n,i} = N_{n,i \max-i}^c \quad (2)$$

where  $i \max$  represents the last position of each sequence.

**Table 3.3:** Anti-Nucleotide Matrices (ANM) generated from NM in Table 3.2

Position	1	2	3	4	5
A	0.33	0.00	0.00	1.00	0.00
T	0.33	0.67	0.00	0.00	1.00
C	0.00	0.33	0.00	0.00	0.00
G	0.33	0.00	1.00	0.00	0.00

Next, the system uses NM and ANM to compute a Degenerate Matrices (DM) and an Anti-Degenerate Matrices (ADM) which will be used to generate a degenerate sequence DM can be computed from Eq. (3) and Table 3.4

$$D_{n,i} = \begin{cases} \frac{1}{g_i}, & N_{n,i} > 0 \\ 0, & N_{n,i} = 0 \end{cases} \quad (3)$$

where  $n$  represents nucleotide base (A,T,C,G),  $D_{n,i}$  is the probability of having nucleotide  $n$  at position  $i$ ,  $g_i$  is the number of nucleotide base at position  $i$  whose  $N_{n,i}$  is greater than 0.

**Table 3.4:** Degenerate Matrices (DM) generated from NM in Table 3.2

Position	1	2	3	4	5
A	1.00	0.00	0.00	0.5	0.33
T	0.00	1.00	0.00	0.00	0.33
C	0.00	0.00	1.00	0.00	0.33
G	0.00	0.00	0.00	0.5	0.00

On the other hand, ADM can be computed from ANM as expressed in Eq. (4) and Table 3.5

$$M_{n,i} = \begin{cases} \frac{1}{p_i}, & A_{n,i} > 0 \\ 0, & A_{n,i} = 0 \end{cases} \quad (4)$$

where  $p_i$  is the number of nucleotide base at position  $i$  whose  $A_{n,i}$  is greater than 0.

Table 3.5: Anti-Degenerate Matrices (ADM) generated from ANM in Table 3.3

Position	1	2	3	4	5
A	0.33	0.00	0.00	1.00	0.00
T	0.33	0.5	0.00	0.00	1.00
C	0.00	0.5	0.00	0.00	0.00
G	0.33	0.00	1.00	0.00	0.00

Then a Consensus Matrices (CM) and an Anti-Consensus Matrices (ACM), which will be used to generate a consensus sequence, are generated from NM and ANM, respectively. CM is the probability of having the most probable nucleotide at each position  $i$ , as expressed in Eq. (5) and Table 3.6

$$C_{n,i} = \begin{cases} \frac{1}{h_i}, & N_{n,i} \geq x_i \\ 0, & N_{n,i} < x_i \end{cases} \quad (5)$$

where  $n$  represents nucleotide base (A,T,C,G),  $C_{n,i}$  is the probability of having consensus nucleotide at position  $i$  for CM,  $x_i$  is the maximum value of  $N_{n,i}$  at position  $i$  and  $h_i$  is the number of nucleotide base with  $N_{n,i}$  that is greater than or equal to  $x$  at position  $i$ .

**Table 3.6:** Consensus Matrices (CM) generated from NM in Table 3.2

Position	1	2	3	4	5
A	1.00	0.00	0.00	1.00	0.33
T	0.00	1.00	0.00	0.00	0.33
C	0.00	0.00	1.00	0.00	0.33
G	0.00	0.00	0.00	0.00	0.00

Next, ACM can be computed from ANM as expressed in Eq. (6) and Table 3.7

$$R_{n,i} = \begin{cases} \frac{1}{q_i}, & A_{n,i} \geq y_i \\ 0, & A_{n,i} < y_i \end{cases} \quad (6)$$

where  $y_i$  is the maximum value of  $A_{n,i}$  at position  $i$  and  $q_i$  is the number of nucleotide base with  $A_{n,i}$  that is greater than or equal to  $y$  at position  $i$ .

**Table 3.7:** Anti-Consensus Matrices (ACM) generated from ANM in Table 3.3

Position	1	2	3	4	5
A	0.33	0.00	0.00	1.00	0.00
T	0.33	1.00	0.00	0.00	1.00
C	0.00	0.00	0.00	0.00	0.00
G	0.33	0.00	1.00	0.00	0.00

Next, a degenerate sequence, DS, can be generated by selecting IUPAC code that covers all possibilities where  $D_{n,i}$  is not equal to 0, whereas an anti-degenerate sequence AS can be generated by selecting IUPAC code that covers all possibilities where  $M_{n,i}$  is greater than 0. A consensus sequence, CS, can be generated by selecting IUPAC code that covers all possibilities where  $C_{n,i}$  is greater than 0 and an anti-consensus sequence, RS, can be generated by selecting IUPAC code that covers all possibilities where  $R_{n,i}$  is greater than 0 as expressed in Table 3.8

Table 3.8: Sequence generated from ANM in Table 3.3

Position	1	2	3	4	5
Seq 1	D	Y	G	A	T

### 3.2. Primer design

First, forward primers are generated from CM and DM based on nearest-neighbor thermodynamic parameters of matched and mismatched nucleotide sequences. The primers are generated from 3' to 5' and the Gibbs Free Energy of duplex formation ( $\Delta G$ ) with ANM is calculated. This  $\Delta G$  is then used to select the best nucleotide for each position of a sequence from either a degenerate sequence or a consensus sequence at that position using our proposed technique called Dynamic Pattern Matching (DPM), as expressed in Eq. (7)-(8).

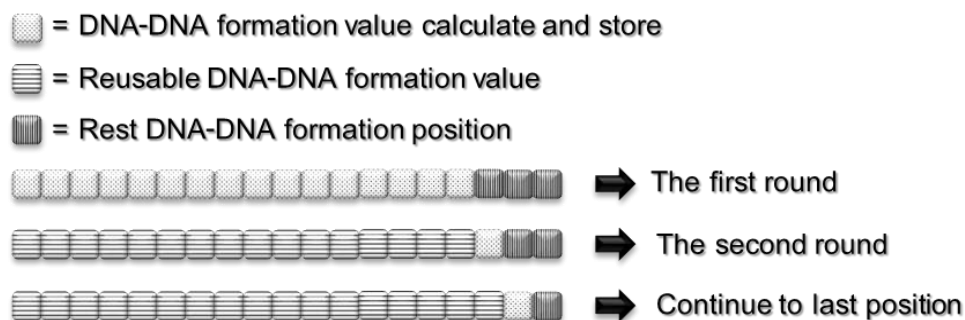
$$z_i = \max(G_{c,i}, G_{d,i}) \quad (7)$$

$$S_i = \begin{cases} CS_i, & CS_i = z_i \\ DS_i, & DS_i = z_i \end{cases} \quad (8)$$

where  $G_{c,i}$  is  $\Delta G$  of consensus nucleotide binding with an antisense strand DNA at position  $i$ ,  $G_{d,i}$  is an average  $\Delta G$  of all possibilities of degenerate nucleotide binding with an antisense strand DNA at position  $i$ ,  $z_i$  is the maximum of the two values,  $CS_i$ ,  $DS_i$ , and  $S_i$  are a consensus sequence, a degenerate sequence, and a primer sequence at position  $i$ , respectively.

By using DPM technique, a user can start generating a primer sequence from 14 nucleotides where all necessary properties of this initial sequence are stored. When a user wants to generate a longer primer sequence, the DPM takes advantage of various coherence properties of a sequence; i.e., the technique only needs to calculate the unknown values for the extended sequence and to reuse previously stored properties such as  $\Delta G$ s of the existing primer sequence without recalculating them. The length of a primer sequence can be in the range of 14 – 30 nucleotides. After the length of the

sequence reaches 30, DPM technique starts to find the next primer sequence by shifting to the next position in CS and DS and repeating the process until the end of ANM.



**Figure 3.2** Dynamic Pattern Matching (DPM) work flow

After primer design process is completed, the primers are ranked according to two main criteria. First,  $\Delta G$ s of the first five sequences of 3' are ranked in increasing order and if there are two or more  $\Delta G$ s with the same value then  $\Delta G$ s of the whole primer sequences are used to rank the sequences. The proposed method can also be used to generate reverse primers from ACM, ADM, and NM.

### 3.3. Property filtering

In this last step, our proposed method provides a user with options to include additional criteria. All primer sets are checked by user's setting properties, such as GC%, 3'GC%, melting temperature, primer length, degeneracy, single repeat and di-repeat. Any primers that meet all users' criteria are returned as results.

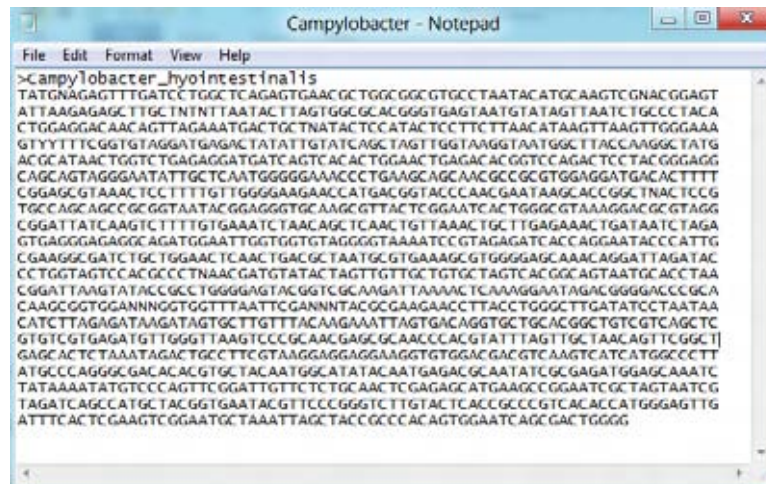


## CHAPTER IV

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1. PCR primer evaluation

In the experiment, the sequences that have been used for test are collected from Genbank database (URL: <http://www.ncbi.nlm.nih.gov/>) and Carbohydrate-Active enZymes Database (CAZy, URL: <http://www.cazy.org/>). The sequences are collected in FASTA format and saved in .txt file by notepad. The example of FASTA format is shown in Figure 4.1.



```
>campylobacter_hyo1intestinalis
TATGNAGAGTTTGATCC TGGC TC AGAGTGAAC GC TGGC GGC GTGCC TAATAC ATGC AAGTC GNACGGAGT
ATTAAAGAGAGC TTGC TNNTTAAATAC TTAGTGGCGCACGGGTGAGTAATGTATAGTTAATC TGCCCTACA
CTGGAGGAC AAC AGTTAGAAATGACTGC TNATACTCC ATACTCC TTC TTAAC ATAAGTTAAGTTGGGAAA
GTYYTTTCGGTGTAGGATGAGACTATAT TGTATCAGC TAGT TGGTAAGGTAAATGGC TTACC AAGGC TATG
ACGCATAAC TGGTC TGAGAGGATGATC ACTC AC ACTGGAAC TGAGAC AC GGTCCAGAC TCCTAC GGGAGG
CAGCAGTAGGGAAATTTGCTCAATGGGGAAACCTGGAAGCAGCAACGCCGC GTGGAGGATGACACTTTT
CGGAGCGTAAAC TCC TTTTGGGGAAAGAACCATGACGGTACCC AAC GAATAAGCACCGGC TNAC TCCG
TGCAGCAGCCGC GGTAATAC GGGGGTGC AAGCGTTACTC GGAATC ACTGGGC GTAAGGAC GC TAGG
CGGAT TATCAAGTCTTTTGTGAAATC TAACAGC TCAACTGT TAAACTGC TTGAGAAAC TGATAATC TAGA
GTGAGGGAGAGGC AGATGGAATTGGTGGTGTAGGGGTAAAATCC GTAGAGATC ACC AGGAATACCC ATTG
CGAAGGC GATC TGC TGGAAC TCAAC TGAC GC TAATGC GTGAAAGC GTGGGAGCAAC AGGAT TAGATAC
CC TGGTAGTCCACGCC TNAAC GATGTATAC TAGTTGTTC TGTGC TAGTCACGCC AGTAATGC ACC TAA
CGGAT TAAGTATACC GCC TGGGGAGTAC GGTCCGCAAGATTAAAAC TCAAGGAAATAGACGGGGACCCGC A
CAAGCCGTTGANNNGTGGTTTAATTCGANNNTACCCGAAGAACC TTACCTGGGCTTGATATCC TAATAA
CATCTTAGAGATAAGATAGTGC TTGTTTACAAGAAATTAAGTGACAGGTGCTGCACGGCTGCTCAGCTC
GTGCTGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCACGATTTAGTTGC TAACAGTTCCGGCT
GAGCACTCAAATAGACTGCC TTC GTAAGGAGGAGGAAGGTGTGGAC GACGTC AAGTC ATC ATGGCCCTT
ATGCCAGGGCGACACACGTGC TACAATGGCATA TACAATGAGACGCAATATCGC GAGATGGAGCAAACT
TATAAAATATGTC CAGTTC GGATTGTTCTGCAAC TC GAGAGC ATGAAAGC GGAATCG TAGTAATCG
TAGATCAGCCATGC TACGGTGAATAC GTTCCCGGGTCTTGTACTC ACCGCCGTCACACCATGGGAGTTG
ATTTAC TC GAAGTCGGAATGC TAAATTAGCTACC GCCCACAGTGGGAATCAGCGACTGGGG
```

Figure 4.1. Example of a sequence in FASTA format

After collection, sequences are aligned by using ClustalX 2.0 [27] to prepare .aln file that is used as input for the proposed system. First, sequences are selected randomly for alignment and after that the sequences that show more similarity are chosen for the next alignment.

The experiment is separated into two parts. In the first part, degenerate primers are designed from 16s rRNA genes of difference genus bacteria and compared with accepted universal primers, as shown in Table 4.1. The second part, degenerate

primers are designed from Glycoside Hydrolase Family 45 and compared with the primer set that has been proven positive in laboratory [28], as shown in **Table 4.2**.

**Table 4.1:** 26 universal primers for comparison results from the proposed system

Primer name	Sequence (5' to 3')	Reference
27F	AGAGTTTGATCMTGGCTCAG	[18]
337F	GACTCCTACGGGAGGCWGCAG	[19]
341F	CCTACGGGAGGCAGCAG	[20]
357F	CTCCTACGGGAGGCAGCAG	[21]
530F	GTGCCAGCMGCCGCGG	[21]
515F	GTGCCAGCMGCCGCGGTAA	[22]
785F	GGATTAGATACCCTGGTA	[23]
928F	TAAACTYAAAKGAATTGACGGG	[21]
968F	AACGCGAAGAACCTTAC	[24]
1100F	YAACGAGCGCAACCC	[19]
1114F	GCAACGAGCGCAACCC	[21]
1406F	TGYACACACCTCCCGT	[21]
336R	ACTGCTGCSYCCCGTAGGAGTCT	[19]
342R	CTGCTGCSYCCCGTAG	[21]
518R	GTATTACCGCGGCTGCTGG	[20]
519R	GWATTACCGCGGCKGCTG	[22]
534R	ATTACCGCGGCTGCTGG	[20]
805R	GACTACCAGGGTATCTAATC	[25]
907R	CCGTCAATTCCTTTRAGTTT	[21]
926R	CCGTCAATTCMTTGGAGTTT	[20]
1100R	GGGTTGCGCTCGTTG	[21]
1392R	ACGGGCGGTGTGTRC	[21]
1401R	CCGTGTGTACAAGACCC	[24]
1492R	TACGGYTACCTTGTTACGACTT	[21]

**Table 4.1 (cont.):** 26 universal primers for comparison results from the proposed system

Primer name	Sequence (5' to 3')	Reference
U1492R	GGTTACCTTGTTACGACTT	[22]
1525R	AAGGAGGTGWTCCARCC	[21]

**Table 4.2:** Degenerate primer for GH 45 [28]

Primer name	Sequence (5' to 3')
Cel45F	ACNMGNTAYTGGGAYTGYTG
Cel45R	AANRYNCCNAVNCCNCCNCCNGG

#### 4.1.1 Bacterial universal primer designing

In this experiment, we separate it into seven sub-experiments. Each experiment contains different amount of sequences and different optional setting values.

##### 4.1.1.1 The first experiment

In the first experiment, we use 44 sequences of different bacteria genus in 16s ribosomal RNA, as shown in **Table 4.3**.

**Table 4.3:** 16s rRNA genes of 44 organisms for system testing (first experiment)

Organism name	Accession number
<i>Acyrtosiphon symbiont</i>	M27040
<i>Azospirillum brasilense</i>	FR745918
<i>Borrelia burgdorferi</i>	GQ478290
<i>Brachyspira hyodysenteriae</i>	U23035
<i>Campylobacter hyointestinalis</i>	M65010
<i>Chlorobium chlorovibrioides</i>	NR044918
<i>Citrobacter freundii</i>	AB210978
<i>Dermacoccus nishinomiyaensis</i>	NR044872

Table 4.3 (cont.): 16s rRNA genes of 44 organisms for system testing (first experiment)

Organism name	Accession number
<i>Desulfosporosinus hippei</i>	NR044919
<i>Enterobacter cloacae</i>	JN644498
<i>Escherichia coli</i>	FJ950694
<i>Fangia hongkongensis</i>	NR041041
<i>Gluconobacter cerinus</i>	NR041048
<i>Hippea maritima</i>	NR044940
<i>Hymenobacter chitinivorans</i>	NR044945
<i>Hyphomicrobium methylovorum</i>	NR026430
<i>Hyphomonas jannaschiana</i>	M83806
<i>Ignicoccus islandicus</i>	NR044910
<i>Klebsiella variicola</i>	JN644499
<i>Leptospira interrogans</i>	DQ840043
<i>Lishizhenia caseinilytica</i>	AB176674
<i>Lysobacter koreensis</i>	NR041014
<i>Macrococcus bovicus</i>	NR044928
<i>Microbulbifer variabilis</i>	NR041021
<i>Myroides pelagicus</i>	NR041042
<i>Myxococcus fulvus</i>	AJ233917
<i>Nannocystis exedens</i>	AJ233946
<i>Neisseria gonorrhoeae</i>	X07714
<i>Nocardia anaemiae</i>	NR041010
<i>Novosphingobium naphthalenivorans</i>	AB684349
<i>Pelotomaculum schinkii</i>	NR044877
<i>Polyangium cellulorum</i>	M94282
<i>Providencia rettgeri</i>	JN644501
<i>Pseudomonas aeruginosa</i>	FN645737

Table 4.3 (cont.): 16s rRNA genes of 44 organisms for system testing (first experiment)

Organism name	Accession number
<i>Rhizobium leguminosarum</i>	HQ218437
<i>Serpulina hyodysenteriae</i>	M57741
<i>Serratia marcescens</i>	AY566180
<i>Shewanella colwelliana</i>	AY653177
<i>Skermanella parooensis</i>	NR044876
<i>Sphingomonas haloaromaticamans</i>	NR044902
<i>Starkeya koreensis</i>	AB166877
<i>Stenotrophomonas maltophilia</i>	JN644502
<i>Thioreductor micantisoli</i>	NR041022
<i>Treponema denticola</i>	AR621358

By using the proposed system with default setting criteria without optional primer properties checking as shown in Figure 4.2, we have found that the system can design 4085 forward primers and 4085 reverse primers. After  $\Delta G$  and 3' $\Delta G$  filtering, most of the primers are removed. After compared with the universal primer in Table 4.1, no universal primers are found, as shown in Table 4.4.

Step 2: Inform your alignment file

Select type of sequences

Amino

Nucleotide

Sequences in file

Step 3: Setting main primer properties

Setup the value (default values are already filled in the box)

- Whole Primer/Template Gibb's energy ( kcal/mol )  -

- 3' Primer/Template Gibb's energy ( kcal/mol )  -

Step 4: Setting optional primer properties

Select properties and setup the value (if left blank check box, that property is not used to consider primer)

GC contents ( % )  -

3' GC contents ( % )  -

Melting Temperature ( °C )  -

Primer length ( bp )  -

Maximum Degeneracy

Maximum Single repeat acceptance ( bp )

Maximum Di-repeat acceptance ( di-bp )

Figure 4.2. The first experiment criteria setting

Table 4.4: The first experimental results

System progress	Number of forward primer	Number of reverse primer
Designing step	4085	4085
$\Delta G$ filtering ( $> -9$ kcal/mol)	1348	1471
$3'\Delta G$ filtering ( $> -5$ kcal/mol)	1095	1282
Universal primer found	0	0

According to the results, universal primer cannot be found, and it is found that  $\Delta G$  and  $3'\Delta G$  filtering caused the number of forward primers and reverse primers to be dramatically decreased at 73.19% and 68.62%, respectively. From raw output of the system, most  $\Delta G$  and  $3'\Delta G$  values of primers are lower than  $-9$  kcal/mol and  $-5$  kcal/mol, respectively. These results indicate that the technique used for calculating  $\Delta G$  may have error. And we have also found 3 sequences that are not similar to the other ones as shown in Figure 4.3.

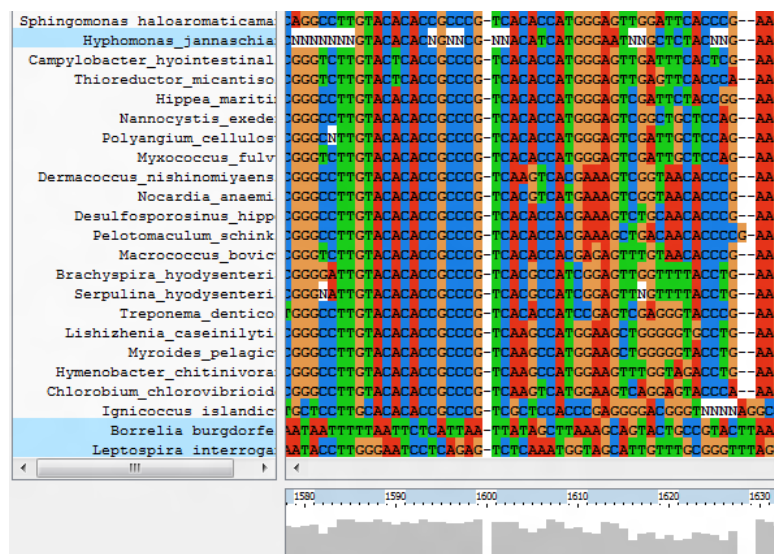


Figure 4.3. The first aligned sequence experiment

According to Figure 4.3, the three highlighted rows are the sequences that are not similar to the other ones (*Borrelia burgdorferi*, *Hyphomonas jannaschiana* and *Leptospira interrogans*). So in the second experiment, these 3 sequences have been removed from the test set and are aligned again.

#### 4.1.1.2 The second experiment

In the second experiment, we use 41 sequences of different bacteria genus in 16s ribosomal RNA, as shown in Table 4.5.

Table 4.5: 16s rRNA genes of 41 organisms for system testing (second experiment)

Organism name	Accession number
<i>Acyrtosiphon symbiont</i>	M27040
<i>Azospirillum brasilense</i>	FR745918
<i>Brachyspira hyodysenteriae</i>	U23035
<i>Campylobacter hyointestinalis</i>	M65010
<i>Chlorobium chlorovibrioides</i>	NR044918
<i>Citrobacter freundii</i>	AB210978
<i>Dermacoccus nishinomiyaensis</i>	NR044872

Table 4.5 (cont.): 16s rRNA genes of 41 organisms for system testing (second experiment)

Organism name	Accession number
<i>Desulfosporosinus hippei</i>	NR044919
<i>Enterobacter cloacae</i>	JN644498
<i>Escherichia coli</i>	FJ950694
<i>Fangia hongkongensis</i>	NR041041
<i>Gluconobacter cerinus</i>	NR041048
<i>Hippea maritime</i>	NR044940
<i>Hymenobacter chitinivorans</i>	NR044945
<i>Hyphomicrobium methylovorum</i>	NR026430
<i>Ignicoccus islandicus</i>	NR044910
<i>Klebsiella variicola</i>	JN644499
<i>Lishizhenia caseinilytica</i>	AB176674
<i>Lysobacter koreensis</i>	NR041014
<i>Macrococcus bovicus</i>	NR044928
<i>Microbulbifer variabilis</i>	NR041021
<i>Myroides pelagicus</i>	NR041042
<i>Myxococcus fulvus</i>	AJ233917
<i>Nannocystis exedens</i>	AJ233946
<i>Neisseria gonorrhoeae</i>	X07714
<i>Nocardia anaemiae</i>	NR041010
<i>Novosphingobium naphthalenivorans</i>	AB684349
<i>Pelotomaculum schinkii</i>	NR044877
<i>Polyangium cellulorum</i>	M94282
<i>Providencia rettgeri</i>	JN644501
<i>Pseudomonas aeruginosa</i>	FN645737
<i>Rhizobium leguminosarum</i>	HQ218437



Table 4.5 (cont.): 16s rRNA genes of 41 organisms for system testing (second experiment)

Organism name	Accession number
<i>Serpulina hyodysenteriae</i>	M57741
<i>Serratia marcescens</i>	AY566180
<i>Shewanella colwelliana</i>	AY653177
<i>Skermanella parooensis</i>	NR044876
<i>Sphingomonas haloaromaticamans</i>	NR044902
<i>Starkeya koreensis</i>	AB166877
<i>Stenotrophomonas maltophilia</i>	JN644502
<i>Thioreductor micantisoli</i>	NR041022
<i>Treponema denticola</i>	AR621358

By using the proposed system with minor changes in setting criteria without optional primer properties checking as shown in **Figure 4.4**, we have found that the system could design 9051 forward primers and 9051 reverse primers. After  $\Delta G$  and  $3'\Delta G$  filtering, some primers have been removed. After comparing with the universal primers in **Table 4.1**, we have found 6 universal primers, as shown in **Table 4.4** and **Table 4.5**.

Step 2: Inform your alignment file . . .

Select type of sequences

Amino

Nucleotide

Sequences in file

Step 3: Setting main primer properties

Setup the value (default values are already filled in the box)

- Whole Primer/Template Gibb's energy ( kcal/mol )  -

- 3' Primer/Template Gibb's energy ( kcal/mol )  -

Step 4: Setting optional primer properties

Select properties and setup the value (if left blank check box, that property is not used to consider primer)

GC contents ( % )  -

3' GC contents ( % )  -

Melting Temperature ( °C )  -

Primer length ( bp )  -

Maximum Degeneracy

Maximum Single repeat acceptance ( bp )

Maximum Di-repeat acceptance ( di-bp )

Figure 4.4. The second experiment criteria setting

Table 4.6: The second experimental results

System progress	Number of forward primer	Number of reverse primer
Designing step	9501	9501
$\Delta G$ filtering ( $> -29$ kcal/mol)	9117	8889
3' $\Delta G$ filtering ( $> -10$ kcal/mol)	8610	8453
Universal primer found	3	3

Table 4.7: Universal primers designed in the second experiment

Primer name	Primer sequence	Reference
341F	CCTACGGGAGGCAGCAG	[31]
1114F	GCAACGAGCGCAACCC	[32]
968F	AACGCGAAGAACCTTAC	[35]
518R	GTATTACCGCGGCTGCTGG	[31]
534R	ATTACCGCGGCTGCTGG	[31]
1100R	GGTTGCGCTCGTTG	[32]

According to the results, after designing step, we have found 132.58% increase in primer output because of three sequences that have less similarity have been removed. Thus,  $\Delta G$  and  $3'\Delta G$  filter are more effective. Only 4.04% forward primers and 6.44% reverse primers have been removed. Six universal primers have been found,  $\Delta G$  and  $3'\Delta G$  of these 6 universal primers are shown in **Table 4.8**.

**Table 4.8:**  $\Delta G$  and  $3'\Delta G$  of universal primer output from the second experiment

Primer name	$\Delta G$ (kcal/mol)	$3'\Delta G$ (kcal/mol)
341F	-21.92	-6.44
1114F	-20.94	-6.48
968F	-16.39	-3.69
518R	-18.87	-5.71
534R	-17.18	-5.71
1100R	-13.88	-4.07

According to **Table 4.8**,  $\Delta G$  and  $3'\Delta G$  are lower than the optimum values [37]. But in this thesis, the thermodynamic values are used for predicting and comparing stability purpose, not for finding the exact value. Thus, we set  $\Delta G$  and  $3'\Delta G$  values as -29 kcal/mol and -10 kcal/mol respectively in the next experiment, and we have found four sequences that are less similar to others (*Rhizobium leguminosarum*, *Pelotomaculum schinkii*, *Hippea maritime* and *Ignicoccus islandicus*). In the third experiment, these four sequences are removed from the test set and are aligned again.

#### 4.1.1.3 The third experiment

In the third experiment, we use 37 sequences of different bacteria genus in 16s ribosomal RNA, as shown in **Table 4.9**.

Table 4.9: 16s rRNA genes of 37 organisms for system testing (third experiment)

Organism name	Accession number
<i>Acyrtosiphon symbiont</i>	M27040
<i>Azospirillum brasilense</i>	FR745918
<i>Brachyspira hyodysenteriae</i>	U23035
<i>Campylobacter hyointestinalis</i>	M65010
<i>Chlorobium chlorovibrioides</i>	NR044918
<i>Citrobacter freundii</i>	AB210978
<i>Dermacoccus nishinomiyaensis</i>	NR044872
<i>Desulfosporosinus hippei</i>	NR044919
<i>Enterobacter cloacae</i>	JN644498
<i>Escherichia coli</i>	FJ950694
<i>Fangia hongkongensis</i>	NR041041
<i>Gluconobacter cerinus</i>	NR041048
<i>Hymenobacter chitinivorans</i>	NR044945
<i>Hyphomicrobium methylovorum</i>	NR026430
<i>Klebsiella variicola</i>	JN644499
<i>Lishizhenia caseinilytica</i>	AB176674
<i>Lysobacter koreensis</i>	NR041014
<i>Macrococcus bovicus</i>	NR044928
<i>Microbulbifer variabilis</i>	NR041021
<i>Myroides pelagicus</i>	NR041042
<i>Myxococcus fulvus</i>	AJ233917
<i>Nannocystis exedens</i>	AJ233946
<i>Neisseria gonorrhoeae</i>	X07714
<i>Nocardia anaemiae</i>	NR041010
<i>Novosphingobium naphthalenivorans</i>	AB684349
<i>Polyangium cellulorum</i>	M94282

Table 4.9 (cont.): 16s rRNA genes of 37 organisms for system testing (third experiment)

Organism name	Accession number
<i>Providencia rettgeri</i>	JN644501
<i>Pseudomonas aeruginosa</i>	FN645737
<i>Serpulina hyodysenteriae</i>	M57741
<i>Serratia marcescens</i>	AY566180
<i>Shewanella colwelliana</i>	AY653177
<i>Skermanella parooensis</i>	NR044876
<i>Sphingomonas haloaromaticamans</i>	NR044902
<i>Starkeya koreensis</i>	AB166877
<i>Stenotrophomonas maltophilia</i>	JN644502
<i>Thioreductor micantisoli</i>	NR041022
<i>Treponema denticola</i>	AR621358

By using the proposed system with minor changes in setting criteria without optional primer properties checking as shown in **Figure 4.5**, we have found that the system could design 10397 forward primers and 10397 reverse primers. After  $\Delta G$  and  $3'\Delta G$  filtering, some primers were been removed. After comparing with the universal primers in **Table 4.1**, we have found seven universal primers, as shown in **Table 4.10** and **Table 4.11**.

Step 2: Inform your alignment file .

Select type of sequences

Amino

Nucleotide

Sequences in file

Step 3: Setting main primer properties

Setup the value (default values are already filled in the box)

- Whole Primer/Template Gibb's energy ( kcal/mol )  -

- 3' Primer/Template Gibb's energy ( kcal/mol )  -

Step 4: Setting optional primer properties

Select properties and setup the value (if left blank check box, that property is not used to consider primer)

GC contents ( % )  -

3' GC contents ( % )  -

Melting Temperature ( °C )  -

Primer length ( bp )  -

Maximum Degeneracy

Maximum Single repeat acceptance ( bp )

Maximum Di-repeat acceptance ( di-bp )

Figure 4.5. The third experiment criteria setting

Table 4.10: The third experimental results

System progress	Number of forward primer	Number of reverse primer
Designing step	10397	10397
$\Delta G$ filtering ( $> -29$ kcal/mol)	9861	9771
3' $\Delta G$ filtering ( $> -10$ kcal/mol)	9248	9326
Universal primer found	4	3

Table 4.11: Universal primers designed in the third experiment

Primer name	Primer sequence	Reference
357F	CTCCTACGGGAGGCAGCAG	[32]
341F	CCTACGGGAGGCAGCAG	[31]
1114F	GCAACGAGCGCAACCC	[32]
968F	AACGCGAAGAACCTTAC	[35]
518R	GTATTACCGCGGCTGCTGG	[31]
534R	ATTACCGCGGCTGCTGG	[31]

Table 4.11 (cont.): Universal primers designed in the third experiment

Primer name	Primer sequence	Reference
1100R	GGGTTGCGCTCGTTG	[32]

According to the results, after designing step, we have found 9.43% increase in primers output comparing with the second experiment because of four sequences that have less similarity have been removed.  $\Delta G$  and 3' $\Delta G$  filter are more effective, only 11.05% forward primers and 10.30% reverse primers have been removed. From aligned sequence as shown in Figure 4.6, we could group the sequences that have more similarity and use them for experiment 4.



Figure 4.6. The third experiment aligned sequence

According to Figure 4.5, we have found 10 sequences that have more similarity to each other. Thus in the fourth experiment, these 10 sequences have been removed from the test set and are aligned again.

## 4.1.1.4 The fourth experiment

In the fourth experiment, we use 10 sequences of different bacteria genus in 16s ribosomal RNA, as shown in **Table 4.12**.

**Table 4.12:** 16s rRNA genes of 10 organisms for system testing (fourth experiment)

Organism name	Accession number
<i>Brachyspira hyodysenteriae</i>	U23035
<i>Chlorobium chlorovibrioides</i>	NR044918
<i>Citrobacter freundii</i>	AB210978
<i>Enterobacter cloacae</i>	JN644498
<i>Escherichia coli</i>	FJ950694
<i>Macrococcus bovicus</i>	NR044928
<i>Providencia rettgeri</i>	JN644501
<i>Pseudomonas aeruginosa</i>	FN645737
<i>Rhizobium leguminosarum</i>	HQ218437
<i>Sphingomonas haloaromaticamans</i>	NR044902

By using the proposed system with minor changes in setting criteria without optional primer properties checking as shown in **Figure 4.7**, we have found that the system could design 17052 forward primers and 17052 reverse primers. After  $\Delta G$  and  $3'\Delta G$  filtering, some primers have been removed. After comparing with the universal primers in **Table 4.1**, we have found seven universal primers, as shown in **Table 4.13** and **Table 4.14**.



Step 2 : Inform your alignment file .

Select type of sequences

Amino

Nucleotide

Sequences in file

Step 3 : Setting main primer properties

Setup the value (default values are already filled in the box)

- Whole Primer/Template Gibb's energy ( kcal/mol )  -

- 3' Primer/Template Gibb's energy ( kcal/mol )  -

Step 4 : Setting optional primer properties

Select properties and setup the value (If left blank check box, that property is not used to consider primer)

GC contents ( % )  -

3' GC contents ( % )  -

Melting Temperature ( °C )  -

Primer length ( bp )  -

Maximum Degeneracy

Maximum Single repeat acceptance ( bp )

Maximum Di-repeat acceptance ( di-bp )

Figure 4.7. The fourth experiment criteria setting

Table 4.13: The fourth experimental results

System progress	Number of forward primer	Number of reverse primer
Designing step	17052	17052
$\Delta G$ filtering ( $> -29$ kcal/mol)	15878	16163
3' $\Delta G$ filtering ( $> -10$ kcal/mol)	14954	15488
Universal primer found	5	4

Table 4.14: Universal primers designed in the fourth experiment

Primer name	Primer sequence	Reference
785F	GGATTAGATACCCTGGTA	[34]
357F	CTCCTACGGGAGGCAGCAG	[32]
341F	CCTACGGGAGGCAGCAG	[31]
1114F	GCAACGAGCGCAACCC	[32]
968F	AACGCGAAGAACCTTAC	[35]

**Table 4.14 (cont.):** Universal primers designed in the fourth experiment

Primer name	Primer sequence	Reference
518R	GTATTACCGCGGCTGCTGG	[31]
534R	ATTACCGCGGCTGCTGG	[31]
1100R	GGGTTGCGCTCGTTG	[32]
805R	GACTACCAGGGTATCTAATC	[36]

According to the results, after designing step, we have found 64.01% increase in primers output comparing with the third experiment because of these ten sequences that have more similarity.  $\Delta G$  and 3' $\Delta G$  filter remove 14.03% forward primers and 10.10% reverse primers.

According to experiment 1-4, we can prove that the system can design the acceptable primers (in term of coverage) in the real PCR experiment. The accuracy of primer designing cannot be tested because it depends on the user criteria setting.

#### 4.1.2 GH45 specific primer designing

In the this experiment, we use five amino acid sequences of species in Glycoside Hydrolase Family 45, as shown in **Table 4.15**.

**Table 4.15:** Amino acid sequence of 5 organisms for system testing

Organism name	Accession number
<i>Cellvibrio japonicus</i>	ACE82688
<i>Fusarium oxysporum</i>	AAA65589
<i>Humicola grisea</i>	AAE55435
<i>Humicola grisea</i>	BAA74957
<i>Humicola insolens</i>	AAE16508

By using the proposed system with minor changes in setting criteria without optional primer properties checking as shown in **Figure 4.8**, we have found that the

system could design 7973 forward primers and 7973 reverse primers. After  $\Delta G$  and  $3'\Delta G$  filtering, some primers have been removed. After comparing with the GH 45 primers in Table 4.2, we have found a forward primer, as shown in Table 4.16.

The screenshot shows a web-based interface for primer design with the following settings:

- Step 2: Inform your alignment file**
  - Select type of sequences:  Amino,  Nucleotide
  - Sequences in file: 5
- Step 3: Setting main primer properties**
  - Setup the value (default values are already filled in the box)
    - Whole Primer/Template Gibb's energy ( kcal/mol ): -29 - 0
    - 3' Primer/Template Gibb's energy ( kcal/mol ): -10 - 0
- Step 4: Setting optional primer properties**
  - Select properties and setup the value (If left blank check box, that property is not used to consider primer)
    - GC contents ( % ): 40 - 60
    - 3' GC contents ( % ): 60 - 100
    - Melting Temperature ( °C ): 50 - 55
    - Primer length ( bp ): 18 - 23
    - Maximum Degeneracy: 1024
    - Maximum Single repeat acceptance ( bp ): 4
    - Maximum Di-repeat acceptance ( di-bp ): 4

Figure 4.8. Specific primer designing experiment criteria setting

Table 4.16: Specific primer designing experimental result

System progress	Number of forward primer	Number of reverse primer
Designing step	7973	7973
$\Delta G$ filtering ( $> -29$ kcal/mol)	3566	3399
$3'\Delta G$ filtering ( $> -10$ kcal/mol)	2651	2458
Specific primer found	Found	Not found

According to the results, we have found the exact forward primer sequence at the first test. For reverse primer results, we cannot find the exact reverse primer sequence because of primer Cel45R (5' AANRYNCCNAVNCCNCCNCCNGG 3') designing and manual method modification, but we can find some primer that almost similar to Cel45R. In this system, the modification method is not included but we can find

some primers that may be used in the real PCR experiment because it can be almost similar to the primer that have already been proved positive in laboratory.

According to experiments 4.1 and 4.2, the system can design the primers that have already been proved in laboratory and a large amount of primers that probably work in laboratory too.

#### 4.2. Computational complexity evaluation

The system processes aligned sequences in three steps: data reformation, primer design, and property filtering.

The first step is data reformation; using normal linear indexing in two dimensional array for calculating probability of nucleotide, running time depends upon the number of sequences and the length of aligned sequences, or  $O(nm)$ , where  $m$  is the number of sequences and  $n$  is the length of each sequence in sequence alignment.

Step two is potentially the most time consuming step. Primer design by using DPM technique starts the first iteration at 3' of NM. Each position of the first iteration is used to calculate the stability with ADM and ACM as the second iteration. Each selected sequences are used to calculate the other properties that use another iteration. Running time tends to have  $O(n^3)$  complexity, where  $n$  is the length of each sequence in sequence alignment.

The last step is to filter a primer with user's setting properties suitable for amplifying specific PCR conditions. Using brute force strategy, this step can be completed in linear time.

## CHAPTER V

### CONCLUSION

Degenerate primers are used to study gene diversity; the designing of degenerate primer becomes a popular issue for scientists in this field. Although many techniques have been proposed and implemented, especially the hybrid primer technique, this technique can optimize degeneracy value for handle the conflict of coverage and specificity [4][13].

This thesis proposes the new technique for designing a simple hybrid degenerate primer by using Dynamic Pattern Matching (DPM) algorithm. According to the experiment results, the tests of universal bacteria primer with DPM technique are suitable for degenerate primer designing system. From the first results, our proposed system can design nine accepted universal primers from a set of difference genus bacterial sequences. From the second experimental results, the tests of degenerate primer design for amino acid sequences, we have found that the proposed system given the specific primer of the endocellulase gene in GH family 45.

It can be concluded that the proposed system can be used to design degenerate primer within polynomial time. In addition, the degenerate primers that are designed through the system can be modified by user by using optional function in this program. Therefore, the objective of developing primer designing system for specific gene biodiversity study is satisfied.

The following tasks are not built into the proposed system. The first one is selection of primer pairs for nested PCR and multiplex PCR. The second one is calculation of stability of secondary formation such as primer dimer, self-priming and hairpin formation. We exclude secondary formation calculation feature because of the error of thermodynamic value calculation that will cause error in this feature too.

There are three advantages of this system compare with others. The first one is Dynamic Pattern Matching (DPM) algorithm that can reduce complexity into polynomial time at  $O(n^3)$ . The second one is mismatch concept with thermodynamic for degenerate primer design that is not included in any system. The last one is flexible styles of user criteria setting that can be applied according to user's experience.

## REFERENCES

1. K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn and H. Erlich. 1986. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. Vol. 51: pp. 263–273.
2. S. Kwok, S. Chang, J. Sninsky and A. Wang. 1994. A guide to the design and use of mismatched and degenerate primers. PCR Methods and Appl. Vol. 3: pp. S39–S47.
3. A. Cornish-Bowden. 1985. IUPAC-IUB symbols for nucleotide nomenclature. Nucleic Acids Res. Vol. 13: pp. 3021-3030.
4. C. Linhart and R. Shamir. 2002. The degenerate primer design problem. Bioinformatics. Vol. 18: pp. S172-S180.
5. T.M. Roes, E.R. Schultz, J.G. Henikoff, S. Pietrokovski, C.M. Mccallum and S. Henikoff. 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res. Vol. 26: pp. 1628-1635.
6. Z. Pan, R. Barry, A. Lipkin and M. Soloviev. 2007. Selection strategy and design of hybrid oligonucleotide primer for RACE-PCR: cloning a family of toxin-like sequence from *Agelena orientalis*. BMC Molecular Biology. Vol 8.
7. T.M. Rose, J.G. Henikoff and S. Henikoff. 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res. Vol. 31: pp. 3763-3766.
8. J.P. Staheli, J.T. Ryan, A.G. Bruce and R. Boyce, T.M. Rose. 2009. Consensus-degenerate hybrid oligonucleotide primers (CODEHOPs) for the detection of novel viruses in non-human primates. Methods. Vol. 49: pp. 32-41.
9. J.P. Staheli, R. Boyce, D. Kovarik and T.M. Rose. 2011. CODEHOP PCR and CODEHOP PCR Primer Design. Methods in Molecular Biology, pp. 57-73. New York: Springer Science+Business Media,
10. R. Giegerich, F. Meyer and C. Schleiermacher. 1996. GeneFisher - Software Support for the Detection of Postulated Genes. Proceedings of the Fourth International

Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA 94025, pp. 68–77.

11. A.L. Lamprecht, Tiziana Margaria, B. Steffen, A. Sczyrba, S. Hartmeier and R. Giegerich. 2008. GeneFisher-P: Variations of GeneFisher as Process in BiojETI. *BMC bioinformatics*. Vol. 9. pp. s13.
12. Z. Pan, R. Barry, A. Lipkin and M. Soloviev. 2007. Selection strategy and the design of hybrid oligonucleotide primers for RACE-PCR: cloning a family of toxin-like sequences from *Agelena orientalis*. *BMC Molecular Biology*. Vol. 8.
13. C. Linhart and R. Shamir. 2005. The Degenerate Primer Design Problem: Theory and Applications. *Journal of computational biology*. Vol. 12: pp. 431-456.
14. M. Bekaert and E.C. Teeling. 2008. UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Res.* Vol. 36: pp. e56.
15. R. Boutros, N. Stokes, M. Bekaert and E.C. Teeling . 2009. UniPrime2: a web service providing easier Universal Primer design. *Nucleic Acids Res.* Vol. 37: pp. W209-W213.
16. O.M. Jabado, G. Palacios, V. Kapoor, J. Hui, N. Renwick, J. Zhai, T. Briesse and W.L. Lipkin. 2006. Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Res.* Vol. 34: pp. 6605-6611.
17. O.J. Marshall. 2004. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*. Vol. 20: pp. 2471-2472.
18. H.S. Najafabadi, A. Saberi, N. Torabi and M. Chamankhah. 2008. MAD-DPD: degenerate primers with maximum amplification specificity. *BioTechniques*. Vol. 44: pp. 519-526.
19. J. Fredslund and M. Lange. 2007. Primique: automatic design of specific PCR primers for each sequence in a family. *BMC Bioinformatics*. Vol. 8.
20. M.D. Gadberry, S.T. Malcomber, A.N. Doust and E.A. Kellogg. 2004. Primaclade a flexible tool to find conserved PCR primer across multiple species. *Bioinformatic*. Vol. 21: pp. 1263-1264.



21. S.N. Jarman . 2004. Amplicon: software for designing PCR primers on aligned DNA sequences. Bioinformatics. Vol. 20: pp. 1644-1645.
22. H.S. Najafabadi, N. Torabi and M. Chamankhah. Designing multiple degenerate primers via consecutive pairwise alignments. BMC Bioinformatics. Vol. 9.
23. C.H. Yang, Member, IAENG, Y.H. Cheng and L.Y. Chuang. 2010. A Natural PCR-RFLP Primer Design for SNP Genotyping Using a Genetic Algorithm. Proceedings of the International Conference of Engineers and Computer Scientists.
24. A.M.K. Brown, K.S. Lo, P. Guelpa, M. Beaudoin, J.D. Rioux, J.C. Tardif, M.S. Phillips and G. Lettre. 2010. Optimus Primer: A PCR enrichment primer design program for next-generation sequencing of human exonic regions. BMC Research Notes. Vol. 3.
25. L.Y. Li, Q. Li, Y.H. Yu, M. Zhong, L. Yang, Q.H. Wu, Y.R. Qiu and S.Q. Luo. A primer design strategy for PCR amplification of GC-rich DNA sequences. Clinical Biochemistry. Vol. 44: pp. 692-698.
26. S. Rozen and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. Vol. 132: pp. 365-386.
27. M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A.Wilm, R. Lopez, J.D. Thompson, T.J. Gibson and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. Vol. 23: pp. 2947-2948.
28. S. Watcharamul. 2005. Microbial communities of Thai rice field soils during rice straw decomposition. Thesis of Doctoral Degree, University of Newcastle upon Tyne.
29. H. Heuer, M. Krsek, P. Baker, K. Smalla and E.M.H. Wellington. 1997. Analysis of actinomycete communities by specific amplification of genes encoding 16s rRNA and gel-electrophoretic separation in denaturing gradients. Applied and Environmental Microbiology. Vol. 63: pp. 3233-3241.
30. E. Ben-Dov, O.H. Shapiro and A. Kushmaro. 2012. 'Next-base' effect on PCR amplification. Environmental Microbiology Reports. Vol. 4: pp. 183-188.
31. G. Muyzer, S. Hottentrager, A. Teske and C. Waver. 1996. Denaturing gradient gel electrophoresis of PCR-amplified 16s rDNA – a new molecular approach to analyse

- the genetic diversity of mixed microbial communities. Molecular Microbial Ecology Manual. Boston: Kluwer Academic Publishers,
32. D.J. Lane. 1991. 16s/23s rRNA sequencing. Nucleic acid techniques in bacteria systematics. Chichester, England: Academic Press,
  33. S. Tuner, K.M. Pryer, V.P.W. Miao and J.D. Palmer. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. Journal of Eukaryotic Microbiology. Vol. 46: pp. 327-338.
  34. H. Zhang, Y.K. Lee, W. Zhang and H.K. Lee. 2006. Culturable actinobacteria from the marine sponge *Hymeniacidon perleve*: isolation and phylogenetic diversity by 16s rRNA gene-RFLP analysis. Antonie van Leeuwenhoek. Vol. 90: pp. 159-169.
  35. U. Nubel, B. Engelen, A. Felske, J. Snaidr, A. Wieshuber, R.I. Amann, W. Ludwig and H. Backhaus. 1996. Sequence heterogeneities of genes encoding 16s rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J. Bacteriol. Vol. 178: pp. 5636-5643.
  36. J.V. Lopez, P.J. McCarthy, K.E. Janda, R. Willoughby and S.A. Pomponi. 1999. Molecular techniques reveal wide phylogenetic diversity of heterotrophic microbes associated with *Discodermia* spp. (Porifera: Demospongiae). Mem. Queensland Museum. Vol. 44: pp. 329-341.
  37. W. Rychlik. 1993. Primer selection and design for polymerase chain reaction. Nucleic Acids Res. pp. 581-588.

## APPENDICES

## Appendix A

### Sequences of 16s rRNA gene of 44 bacteria species

These are 44 nucleic acid sequences of 16s rRNA gene of bacteria. All are listed in FASTA format. Accession number of each species id located at the end of each species name.

```
>Acyrrhosiphon_symbiont_M27040
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAAGCCTAACACATGCAAGTCGAGCGGC
AGCGAGAAGAGAGCTTGCTCTCTTTGTGCGCAAGCGGCAAACGGGTGAGTAATATCTGGGGATCTACCCA
AAAGAGGGGGATAACTACTAGAAATGGTAGCTAATACCGCATAATGTTGAAAAACCAAAGTGGGGGACCT
TTTGGCCTCATGCTTTTGATGAACCCAGACGAGATTAGCTTGTGGTAGAGTAATAGCCTACCAAGGCA
ACGATCTCTAGCTGGTCTGAGAGGATAACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGG
AGGCAGCAGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCTATGCCGCGTGTATGAAGAAGGC
CTTAGGGTTGTAAAGTACTTTTCAGCGGGGAGGAAAAAAATAAACTAATAATTTTATTTTCGTGACGTTAC
CCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGAGGGTGAAGCGTTAATCAG
AATTACTGGGCGTAAAGAGCGCGTAGGTGGTTTTTTAAGTCAGGTGTGAAATCCCTAGGCTCAACCTAGG
AACTGCATTTGAACTGGAAAAGTAGAGTTTCGTAGAGGGAGGTAGAATTCTAGGTGTAGCGGTGAAATG
CGTAGATATCTGGAGGAATACCCGTGGCGAAAGCGGCTCCTAAACGAAAAGTACACTGAGGCGCGAAA
GCGTGGGGAGCAAACAGGATTAGATAACCCTGGTAGTCCATGCCGTAACGATGTCGACTTGGAGGTGTT
TCCAAGAGAAGTGACTTCCGAAGCTAACGCATTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGCTAAA
ACTCAAATGAATTGACGGGGGCGCACAAGCGGTGGAGCATGTGGTTTTAATTCGATGCAACGCCGAAAAAC
CTTACCTGGTCTTGACATCCACAGAATTCTTTAGAAATAAAGAAGTGCCTTCGGGAGCTGTGAGACAGGT
GCTGCATGGCTGTGTCAGCTCGTGTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCATTATC
CCCTGTTGCCAGCGGTTCCGGCCGGAAGTACAGGAGACTGCCGTTATAAACCGGAGGAAGTGGGGAC
GACGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGTTTTATACAAGAGAAGC
AAATCTGCAAAGACAAGCAAACCTCATAAAGTAAATCGTAGTCCGGACTGGAGTCTGCAACTCGACTCCA
CGAAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCTTGTACACAC
CGCCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGCAGGTATCCTAACCCCTTTAAAAGGAAGGCGCTTA
CCACTTTGTGATTTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCAC
CTCCTTA
```

```
>Azospirillum_brasilense_FR745918
AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCATGCCTAACACATGCAAGTCGAACGAAGGCTTCG
GCCTTAGTGGCGCACGGGTGAGTAACACAGTGGGAACCTGCCTTTCGGTTCGGGATAACGCTCTGAAACGG
ACGCTAACACCGGATACGTCTCTTCGGGAGAAAGTTTACGCCGAGAGAGGGGCCCGCTCCGATTAGGTAG
TTGGTGGGGTAATGGCCACCAAGCCGACGATCGGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGA
CTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATCGGACAATGGGGCAACCCGTGATC
CAGCAATGCCCGTGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTCGCACGCGACGATGATGACGGT
AGCGTGAGAAGAAGCCCGGCTAACTTCGTGCCAGCAGCCGCGTAATACGAAGGGGCGAGCGTTGTTT
GGAATTACTGGGCGTAAAGGGCGCGTAGGGCGCCTGTTAGTCAGAAGTGAAAGCCCCGGGCTTAACCTG
GGAACGGCTTTTGATACTGGCAGGCTTGAGTTCCGGAGAGGATGGTGGAAATCCAGTGTAGAGGTGAAA
TTCTGTAGATATTGGGAAGAACACCGGTGGCGAAGGCGGCCATCTGGACGGACACTGACGCTGAGGCGCGA
AAGCGTGGGGAGCAAACAGGATTAGATAACCCTGGTAGTCCACGCCGTAACGATGAATGCTAGACGCTGG
GGTGCATGCACCTTCGGTGTGCGCCGCTAACGCATTAAGCATTCCGCTGGGGAGTACGGCCGCAAGGTAA
AACTCAAAGGAATTGACGGGGGCCGCAACAAGCGGTGGAGCATGTGGTTTTAATTCGAAGCAACGCGCAGA
ACCTTACCAACCCCTTGACATGTCCATTGCCGGTCCGAGAGATTGGACCTTCAGTTCGGCTGGATGGAACA
CAGGTGCTGCATGGCTGTGTCAGCTCGTGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCC
CTACCGCCAGTTGCCATCATTAGTTGGGCACCTCTGGTGGAACTGCCGGTGAACAAGCCGAGGAAGGCGG
GGATGACGTCAAGTCTCATGGCCCTTATGGGTTGGGCTACACACGTGCTACAATGGCGGTGACAGTGGG
ATGCGAAGTCGCAAGATGGAGCCAATCCCCAAAAGCCGTCTCAGTTCGGATTGCACTCTGCAACTCGGGT
GCATGAAGTTGGAATCGCTAGTAATCGCGGATCAGCACGCCGCGGTGAATACGTTCCCGGGCTTGTACA
CACCGCCGTCACACCATGGGAGTTGGCTTTACCCGAAGGTGGTGCCTAACCGGCAACGGAGGCAGCCA
ACCACGGTCAGGTCAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGT
```

>Borrelia\_burgdorferi\_GQ478290

TCTTTTCTTTTGATAAAAAGAGTTTTTAAAGTCAGTGTATTTTTTTAAGGTTAATAGAAATTTATGTATT  
 AAGCTTTTTTATTAATGATATTAAGCGCTCTATTAATTTTAGATCATTTTTGGGGTTTTAGCTCAGTTGGC  
 TAGAGCATCGGCTTTGCAAGCCGAGGGTCAAGGGTTCGAGTCCCTTAACCTCCATTTGGCATGTGTCTTA  
 RATTGTGATTAAGCATGTTTTCAAGAACTTTGTTAAAGTAATTAATTAAGATGAAACAGAGGAAGTTAG  
 AATTTCTAGGGTAAGGGTGAATCTTTTAATGCTAAGAAGATGTCTAGGAGTAAAAGCAAACCCTTTAT  
 AAAAAATTTGCGTTGTACTAAATATAAGGATTAACAGGATTGCATTTTCCAGTATCCTATTTCTATAG  
 ACGACCTACATTAATTCCTAAATAAAAAGTCCATAATGAATGTGGATTGTTAAGCGTGATGTCTGGGTGGC  
 AAGTTTTGTAACGAGTCGCAACCCTTGTGTGCTTACCACCAGTAAACGACAGGGATTTGCATAAGTCTT  
 TAGGTTATAAGTAGAAGGGTAAGGGTGTGCTGAGTCGTTCTTTTATGTTCTAGGGTTACGCGCTTGC  
 TATAATGGCCTTTACAAAGCTAAGAGAAATAGTGATGTGAATCAAAGCTTAAATCGGGTCTCAGCATGG  
 ATTGCATTTTGAAATTCGACTTTATGAAGTTGTAATCGCTAGTAATCGTATATCGGAATGATACAGTGAA  
 TCGGTTCTCAGTCCCTGTACACACTCGCTCGTTACATCGTCCGAGTATGGGGGATGCCGAAGTATTAT  
 TCTAACCCATAAGGGAGGAAGGTATTTAAGGTATGTTAGGGAAGGGTTGAAGTAGTAATAAGACGAT  
 AGTACTTTCCAGCATGGCTAGATTGGTCTCTTTTAAATAGAAAAGATAAATTAAGGCTAATTTTATGATT  
 ATTCTTTGTGCTTTTATTTTATTAAAGTTTTTTTAAATCAAGGTAATTTTTTAGTTTAAAATTTAGAA  
 AAAATAAACCTTTAGTAAGTGTTTAGTAGTATTTTAGCTTCAGCTTCATATTTAATTAAGCTTTGTTTGA  
 AGTAGAATTTTAGATTTTAGGGTGAAGGGTTTTTATATTATAAATTTTAAATCCAATGGAACG  
 AGAATTATCTCTATTAATAAATTAATAATACATATTGTAATGCAAGTAGGTCTCAATATAGAGATTGAATA  
 ATTTTTAATTTCTCATTAATTATAGCTTAAAGCAGTACTGCCGTACTTAAAACAACGAAATGCATTGGGAC  
 CAGGATGAGTTGAACATCCGACCTCAGGTTTTATCAGACAAA

>Brachyspira\_hyodysenteriae\_U23035

CGTTGGCGATCGCTTAAAGCATGCAAGTCGAGCGGGCTTATTCGGGCAACTGGATAAGTTAGCGGGCAA  
 CTGGTGAGTAACACGTAGGTAATCTGCCGTAGAGTGGGGATAACCCATGGAAACATGGACTAATACCGC  
 ATATACTCTTGCTACATAAAGTAGAGTAGAGGAAAGGAGCAATCCGCTTTACGATGAGCCTGCGGCTATT  
 AGCCTGTTGGTGGGGTAACGGCCTACCAAAGCTACGATAGGTAGCCGACCTGAGAGGGTGACCGGCCACA  
 TTGGGACTGAGATACGGCCAGACTCCTACGGGAGGCAGCAGCTGAGAATCTCCACAATGGACGAAAGT  
 CTGATGGAGCGACATCGCGTGAGGGATGAAGGCCTTCGGGTGTAAACCTCGGAAATTTATCGAAGAATGA  
 GTGACAGTAGATAATGTAAGCCTCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGAGCAAACG  
 TTGCTCGGATTTACTGGGCGTAAAGGGTGAGTAGGCGACTTATAAGTCTAAGGTGAAAGACCGAAGCTC  
 AACTTCGGAACGCCCTCGGATACTGTAAGTCTTGGATATTGTAGGGGATGATGGAATTTCTCGGTGTAGCG  
 GTGGAATGCGCAGATATCGAGAGGAACACCTATAGCGAAGGCAGNCATCTGGGCATTTATCGACGCTGAA  
 TCACGAAAGCTAGGGGAGCAAACAGGCTTAGATACCTTGTAAGTCTTAGCCGTAACCGTTGTACACTAGG  
 TGCTTCTATTTAAATAGGAGTGCCGTAGCTAACGCTTAAAGTGTACCCGCTGAGGATGATGCCGCAAGG  
 GTGAAACTCAAAGAAATTTGACGGGTCCCCGCACAAGTGGTGGAGCATGTGGTTTTAATTCGATGATACGCG  
 AAAAACCTTACCTGGGTTTGAATTGTAAGATGAATGATTTAGAGATAAGTCAGACCGCAAGGACGTTTTTA  
 CATAGGTGCTGCATGGCTGTGCTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCAGCAACGAGCGCAAC  
 CCTCACCTTTGTTGCTACCGAGTAATGTCGGGCACCTTATAGGGGACTGCCTACGTTCAAGTAGGAGGAA  
 GGTGGGGATGATGTCAGTCTCATGGCCCTTATGTCCAGGGCTACACACGTGCTACAATGGCAAGTACA  
 AAGAGAAGCAAGACCGCGAGGTGGAGCAAAACTCAAAAAGTTGCTCAGTTCCGATTGGAGTCTGAAAC  
 TCGACTCCATGAAGTTGGAATCACTAGTAATCGTAGATCAGAACGCTACGGTGAATACGTTCCCGGGGAT  
 TGTACACACCGCCGTCACGCCATCGGAGTTGGTTTTACTGAAAGTCGTTAGCCTAACCGCAAGGGGGG  
 GCGCCGAAGTGGGACTGATGATGAGGGTGAAGTCGTAACAAGGCAGNCGTACCGGAAGGTGTG

>Campylobacter\_hyointestinalis\_M65010

TATGNAGAGTTTGATCCTGGCTCAGAGTGAACGCTGGCGGCTGCCTAATACATGCAAGTCGNACGGAGT  
 ATTAAGAGAGCTTGCTNTNTAATACTTAGTGGCGCACGGGTGAGTAATGTATAGTTAATCTGCCCTACA  
 CTGGAGGACAACAGTTAGAAATGACTGCTNATACTCCATACTCCTTCTTAACATAAGTTAAGTTGGGAAA  
 GTYYTTTCCGTTGATGAGACTATATTGTATCAGCTAGTTGGTAAGGTAATGGCTTACCAGGCTATG  
 ACGCATAACTGGTCTGAGAGGATGATCAGTCACACTGGAAGTCTGAGACACGGTCCAGACTCCTACGGGAGG  
 CAGCAGTAGGGAATATTGCTCAATGGGGGAAACCTGAAGCAGCAACGCCGCGTGGAGGATGACACTTTT  
 CGGACGCTAAACTCCTTTTGTGGGGGAAGAACCCTGACGGTACCCAACGAATAAGCACCGGCTNACTCCG  
 TGCCAGCAGTAACTGTAATACGGAGGGTGAAGCGTTACTCGGAATCACTGGGCGTAAAGGACGCTAGG  
 CGGATTATCAAGTCTTTTGTGAAATCTAACAGCTCAACTGTTAAACTGCTTGAGAAACTGATAATCTAGA  
 GTGAGGGAGAGGCAGATGGAATTGGTGGTGTAGGGGTAATAATCCGTAGAGATCACCAGGAATACCCATTG  
 CGAAGGCGATCTGCTGGAAGTCAACTGACGCTAATGCGTGAAAGCGTGGGGAGCAACAGGATTAGATAC  
 CCTGGTAGTCCACGCCCTNAACGATGTATACTAGTTGTTGCTGTGCTAGTCACGGCAGTAATGCACCTAA  
 CGGATTAAGTATACCGCTGGGGAGTACGGTTCGCAAGATTAATAACTCAAAGGAATAGACGGGGACCCGCA  
 CAAGCGGTGGANNNGTGGTTAATTCGANNTACCGGAAGAACCTTACCTGGGCTTGATATCCTAATAA  
 CATCTTAGAGATAAGATAGTGCTTGTTTACAAGAAATTAGTGACAGGTGCTGCACGGCTGTCTCAGCTC

GTGTCGTGAGATGTTGGGTTAAGTCCCAGCAACGAGCGCAACCCACGTATTTAGTTGCTAACAGTTCGGCT  
 GAGCACTCTAAATAGACTGCCTTCGTAAGGAGGAGGAAGGTGTGGACGACGTCAAGTCATCATGGCCCTT  
 ATGCCAGGGCGACACACGTGCTACAATGGCATATAACAATGAGACGCAATATCGCGAGATGGAGCAAATC  
 TATAAAATATGTCCCAGTTCGGATTGTTCTCTGCAACTCGAGAGCATGAAGCCGGAATCGCTAGTAATCG  
 TAGATCAGCCATGCTACGGTGAATACGTTCCCGGTCTTGTACTCACCGCCCGTCACACCATGGGAGTTG  
 ATTTCACTCGAAGTCGGAATGCTAAATTAGCTACCGCCACAGTGGAATCAGCGACTGGGG

>*Chlorobium\_chlorovibrioides*\_NR044918

AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCATAACACATGCAAGTCAAAGGATAGTTTNT  
 TCGGANACAAGTACTTGGCGCAAGGGTGTAGTAAGGCATANGTAATCTGCCCTTTGGACTGGGATAACCC  
 GAGAAATCGGGGACAATACCAGATGATGCAGCGGAACCGCATGGTTATGTTGTTAAATGATTTATCGCCA  
 AAGGATGAGCCTATGTTCCATCAGGTAGTTGGTAAGGTAACGGCTTACCAAGCCAACGACGGATAGGTGG  
 TCTGAGAGGATGATCAGCCACATTGGAAGTGTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGAAG  
 ATATTGCGCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGGATGATGAAGTCTTCGGAATGTA  
 TCCTTTTGTGGGGAAGAATAGGTGCGCTTGGCGACTCTGACGGTACCCGGCGAATAAGCCACGGCTAAC  
 TCTGTGCCAGCAGCGCGGTGATACAGGGGTGGCAAGCGTGTCCGGATTTACTGGGTGTAAGGGTGGC  
 CAGGCGGACTGATAAGTTCGGGGTTAAATCCATGTGCTCAACACATGCACGGCTTCCGATACTGTGAGTC  
 TTGAGTCTCGAAGAGGAAGATGGAATTTCCGGTGTACGGTGGAAATGTGTAGATATCGGAAAGAACCA  
 GTGGCGAAGGCAGTCTTCTGGTCGAGTACTGACGCTCAGGCACGAAAGCGTGGGAGCAAACAGGATTAG  
 ATACCCTGGTAGTCCACGCCGTAACGATGAATACTAGATGTTGGTCATATTGATCAGTGTGCGAGCTAA  
 CGCATTAAAGTATTCACCTGGGAAGTACGCCCGCAAGGGTGAAGTCCAAAGAATTGACGGGGCCCCGCA  
 CAAACGGTGGATCATGTGGTTAATTCATGCAACNCGAAGAACCTTACCTAGGCTTGAATGTTAGCTA  
 AAGCTCCTGAAAGGGAGTGTCTTCCGGGANTTAGCACAGGTGCTGCATGGCTGTGCTCAGCTCGTGTGCG  
 TGAGATGTTGGGTTAAGTCCCNACGAGCGCAACCCGTACAATTAGTTANTAACAGGTTAAGNTGAGGA  
 CTCTAATTGAAGTGCCTACGCAAGTANAGAGGAAGGAGGGGATGACGTCAGTCAAGTCAAGTGCCTTACGC  
 NTAGGGCCACACCGTATACAATGGCGACTACANAGGGCAAACCCNCGAGGNAANGGAAATCCCTTAAA  
 AGTCGTCTCAGTCCGGATNGGAGTNTGCAACTCGACTCCGTGAAGTTGGAATCGCTAGTAATCGCAGATC  
 ANCATGCTGCGGTGAATGTGTTCCCGGGCCTTGTACACACCGCCCGTCAAGTCATGGAAGTCAGGAGTAC  
 CCAAAGACGCTCGCAGCTTAAAGGTAAGACTGGTAANTGGGACTATGTCTTANCTAGGTNNCCGTA

>*Citrobacter\_freundii*\_AB210978

TTAGTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCCTAACACATGCAAGTCAACGGTAGCACAGA  
 GGAGCTTGCTCCTTGGGTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCCGATGGAGGGGG  
 ATAACACTGGAAACGGTAGCTAATACCGCATAACGTGCGAAGACCAAAGAGGGGGACCTTCGGGCCCTCT  
 TGCCATCGGATGCCCAGATGGGATTAGCTAGTGGGGTAAACGGCTCACCTAGGCGACGATCCCTTA  
 GCTGGTCTGAGAGGATGACCAGCCACACTGGAAGTGTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGT  
 GGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCCTTCGGGTTG  
 TAAAGTACTTTACGCGAGGAGGAAGGCGTTGTGGTTAATAACCGCAGCGATTGACGTTACTCGCAGAAGA  
 AGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGAAGCGTTAATCGGAATTAAGTGGG  
 CGTAAAGCGCACGCAGGCGGTCTGTCAAGTCCGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCC  
 GAAACTGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGAAATCCAGGTGTAGCGGTGAAATGCGTAGAGATC  
 TGGAGGAATACCGGTGGCGAAGGCGGCCCTGGACAAAGACTGACGCTCAGGTGCCAAAGCGTGGGGAG  
 CAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAACGATGTGACTTGGAGGTTGTGCCCTTGGAGG  
 GTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAACTCAAATGA  
 ATTGACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTAATTCGATGCAACGCGAAGAACCCTTACCTAC  
 TCTTGACATCCAGAGAACTTAGCAGAGATGCTTTGGTGCCTTCGGGAACTCTGAGACAGGTGCTGCATGG  
 CTGTGCTCAGCTCGTGTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTTATCCTTTGTTGC  
 CAGCGGTTCCGGCCGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCAAG  
 TCATCATGGCCCTTACGAGTAGGGCTACACACGTGCTACAATGGCATATACAAAGAGAAGCGACCTCGCG  
 AGAGCAAGCGGACCTCATAAAGTATGTGCTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGG  
 AATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCA  
 CACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACTTCGGGAGGGCGCTTACCACCTTTGTGATT  
 CATGACTGGGGGAGGGTGTGTAACAAGGAACCGGG

>*Dermacoccus\_nishinomiyaensis*\_NR044872

CTGGCGGCGTGTAAACACATGCAAGTCAACGATGAAGCACCAGCTTGTGGTGTGGATTAGTGGCGAA  
 CGGGTGTAGTAACACGTGAGTAACCTGCCCTCCTCTGGGATAAGCCTGGGAAACTGGGTCTAATACTGG  
 ATACGACCGATTTCCGCATGGAGTGTGGTGGAAAGTTTTTGTGGTGGGGGATGGACTCGCGCCCTATCA  
 GCTTGTGGTGGGTTAATGGCCTACCAAGCCGACGACGGTAGCCGGCCTGAGAGGGCGACCGCCACAC  
 TGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGAAAGCC  
 TGATGCAGCGACGCCGCTGAGGGATGACGGCCTTCGGGTTGTAACCTCTTTCACCAGGGACGAAGCTA

ACGTGACGGTACCTGGAGAAGAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCG  
 AGCGTTGTCCGGAAATATTGGGGCGTAAAGAGCTTGTAGGCGGTCTGTGCGGTCTGCTGTGAAAGACCGGG  
 GCTTAACTCCGGTTCTGCAGTGGGTACGGGCAGACTAGAGTGTGGTAGGGGAGACTGGAATTCCTGGTGT  
 AGCGGTGAAATGCGCAGATATCAGGAGGAACACCGATGGCGAAGGCAGGTCTCTGGGCCATTACTGACGC  
 TGAGAAGCGAAAGCATGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCATGCCGTAAACGTGGGGCGC  
 TAGGTGTGGGACTCATTCACAGAGTCCGTGCCGAGCTAACGCATTAAGCGCCCCGCCCTGGGGAGTACG  
 GCCGCAAGGCTAAAACCTCAAAGGAATTGACGGGGGCCCGCACAAAGCGCGGAGCATGCGGATTAATTGCA  
 TGCAACGCGAAGAACCTTACCAAGGCTTGACATACACCGGAATCATGCAGAGATGTGTGCGTCTTCGGAC  
 TGGTGTACAGGTGGTGCATGGTTGTGTCAGCTCGTGTGAGATGTTGGGTTAAGTCCCGCAACGAGC  
 GCAACCCCTCGTTCATGTTGCCAGCACTTCGGGTGGGGACTCATGGGAGACTGCCGGGGTCAACTCGGAG  
 GAAGGTGGGGATGACGTCAAATCATCATGCCCTTATGTCTTGGGCTTCACGCATGCTACAATGGCCGGT  
 ACAGAGGGTTGCGAAACCGTGAGGTGGAGCTAATCCCAAAAAACCGGTCTCAGTTCGGATTGGGGTCTGC  
 AACTCGACCCCATGAAGTCGGAGTCGCTAGTAATCGCAGATCAGCAACGCTGCGGTGAATACGTTCCCGG  
 GCCTTGTACACACCGCCCGTCAAGTCACGAAAGTCGGTAAACCCGAAGCCGGTGGCCTAACCCCTGTGG  
 NGGGAGCCGTGCAAGGTGGGACTGGCGATTGGGACTAAGTCGTAACAAGGTAGCCGTACCGGAA

>Desulfosporosinus\_hippeii\_NR044919

AACACATGCAAGTCGAACGCTTAGTGGCGGACGGGTGAGTAAACGCGTGGGTAACCTACCCATAAAATCCGG  
 GACAACCCCTTGGAAACGAGGGCTAATACCGGATAATCTTCGAGCTTGGCATCAAGCTTGAAGGAAAGATG  
 GCCTCTGAATATGCTATCGATTATGGATGGACCCGCGTCTGATTAGCTAGTTGGTGGGGTAAAGCCCTAC  
 CAAGGCGACGATCAGTAGCCGGCCTGAGAGGGTGAACGGCCACACTGGGACTGAGACACGGCCAGACTC  
 CTACGGGAGGCAGCAGTGGGGAATCTTCCGCAATGGACGAAAAGTCTGACGGAGCAACGCCCGGTGATGA  
 TGAAGGTCTTCGGATTGTAAGTACTGTCTTGGGGGAAGAACGGTGCATTTGAAAATATTGAGTCGACAT  
 GACGGTACCCAAGGAGGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGGGCAAGCG  
 TTGTCCGGAATTAATGGGCGTAAAGGGCGCGTAGGGGATTAATTAAGTCTGGTGTGAAAGATCAGGGCTC  
 AACCCCTGAGAGTGCATCGGAAACTGGTATCTTGTGGACAGGAGAGGAAAGTGAATTCACCGTGTAGCG  
 GTGAAATGCGTAGATATGTGGAGGAACACCGGTGTGGAAGACGACTTTCTGGACTGTAACGTGACGCTGAG  
 GCGCGAAAGCGTGGGGAGCACACAGGATTAGATACCCCTGGTAGTCCACGCCGTAACGATGAGTGCTAGG  
 TGTAGAGGGTATCGACCCCTTCTGTGCCGAGTTAACACAATAAGCACTCCGCCCTGGGGAGTACGGCCGC  
 AAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAAATTCGACGCAA  
 CGCGAAGAACCTTACCAAGGCTTGACATCCTACGAATCCATAGGAAACTAGGGAGTGCCCTTCGGGGAGC  
 GTAGAGACAGGTGGTGCATGGTTGTGTCAGCTCGTGTGAGATGTTGGGTTAAGTCCCGCAACGAGC  
 GCAACCCCTGTATTTAGTTGTAAACGCGTAATGGTGGAGCACTCTAGATAGACTGCCGGTGATAAAACCGGA  
 GGAAGGTGGGGATGACGTCAAATCATCATGCCCTTATGTCTTGGGCTACACAGTGTCTACGATGGCCGG  
 TACAGACGGAAGCGAAGCCGCGAGGTGAAGCAAATCCGAGAAAGCCGGTCTCAGTTCGGATTGACAGCTG  
 CAACTCGCCTGCATGAAGTCGGAATCGCTAGTAATCGCAGGTGAGCATACTGCGGTGAATACGTTCCCGG  
 GCCTTGTACACACCGCCCGTACACCACGAAAGTCTGCAACACCCGAAGCCGGTGGAGTAAACCGTAAGG  
 GAGCTAGCCGTGCAAGGTGGGGCCGATAATTGGGGTGAAGTCGTAACAAGGTGGCCGTATCGGAAGGTG  
 GGCTGGATCACCTCCTTTCGGG

>Enterobacter\_cloacae\_JN644498

TTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAGCACAGAGA  
 GCTTGTCTCCGGTGACGAGTGGCGGACGGGTGAGTAATGTCGGAAGTGCCTGATGGAGGGGGATAA  
 CTACTGGAACGGTAGCTAATACCGCATAACGTGCGAAGACCAAAGAGGGGGACCTTCGGGCCCTTTCG  
 ATCAGATGTGCCGATGGGATTAGCTAGTATGTGGGGTAAACGGCTCACCTAGGCGACGATCCCTAGCTG  
 GTCTGAGAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGG  
 AATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAA  
 GTACTTTCAGCGGGGAGGAAGGTGTTGTGGTTAATAACCACAGCAATTGACGTTACCCGCAAGAAGCA  
 CCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGAAGCGTTAATCGGAATTACTGGGCGTA  
 AAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAAGTGCATTCGAAA  
 CTGGCAGGCTGGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGA  
 GGAATACCGGTGGCGAAGGCGGCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAA  
 CAGTATTAGATACCCCTGGTAGTCCACGCCGTAACAGCATGTGATTTGGAGGTTGTGCCCTTGAGGCGTGG  
 CTTCCGGAGCTAACCGGTTAAATCGACCCGAGGGAGTACGGCCGCAAGGTTAAAACCTCAAATGAATG  
 ACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAAATTCGATGCAACGCGAAGAACCTTACCTGGTCTT  
 GACATCCACAGAACTTTCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAGGTGCTGCATGGCTGT  
 CGTCAGCTCGTGTGTAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTTATCCTTTGTGCCAGC  
 GGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCAAGTCAT  
 CATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCCATACAAAGAGAAGCGACCTCGCGAGAG  
 CAAGCGGACCTCATAAAGTGCCTGTCGTCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATC  
 GCTAGTAATCGTAGATCAGAATGCTACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTACACC

ATGGGAGTGGGTTGCAAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACCTTTGTGATTCATG  
ACTGGGGGTGAAGTCGTAACAT

>Escherichia\_coli\_FJ950694

CGCCCTGATTGACGGCTATACACATGCAAGTCGAACGGTAACAGGAAACAGCTTGCTTCTTTGCTGACGA  
GTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTA  
ATACCGCATAACGTGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCAGATGG  
GATTAGCTAGTAGGTGGGGTAACGGCTCCATCCCTAGCGAGCCGAATCCTTAGCCTGGTCTGAGAGGAA  
TGACCAGCCACACTGGGACTGAGAACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCAC  
AATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTTACG  
CGGGGAGGAAGGGAGTAAAGTTAATACCCTTTGCTCATTGACGTTACCCGCGAGAAGAAGCACCGGCTAAC  
TCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTAAGTGGGCGTAAAGCGCACG  
CAGGCGGTTTGTAAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGC  
TTGAGTCTCGTAGAGGGGGTAGAATTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCG  
GTGGCGAAGGCGGCCCTTGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAG  
ATACCCTGGTAGTCCACGCCGTAACGATGTCGACTTGAGGTTGTGCCCTTGAGGCGTGGATTCGGGAG  
CTAACCGGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAACTCAAATGAATTGACGGGGGCC  
CGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGTCTTGACATCCAC  
GGGAAGTTTTTACAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTGCTGCATGGCTGTCGTAGCT  
CGTGTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTATCCTTTGTTGCCAGCGGTCGGGC  
CGGAACTCAAAGGAGACTGCCAGTGATAAATGGAGGAAGGTGGGGATGACGTCAGGTCATCATGGCC  
CTTACGAACCAGGGCTACACACGTGCCTACAATGGACGCATCCAAAGAGAGAGCGAACCCTGCCCGCGAG  
AGCAAGCGGACCTCATAAAGTGCCTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAA  
TCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACA  
CCATGGGAGTGGGTTGCAAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACCTTTGGATGCGA  
GG

>Fangia\_hongkongensis\_NR041041

AGAGTTTGATCCCTGGCTCAGATTGAACGCTGGTGGTATGCTTAACACATGCAAGTCGAACGGACTTGTT  
AACCTGCTTGACGGTTAACGGTTAGTGGCGGACGGGTGAGTAATGCATAGGAATCTGGCTTATGTTGGGG  
GACAACAGTTGGAAACGGCTGCTAATACCGCATATTTCTAATGATGAAAGGTGCCTTAGGGTGTGCCA  
TGAGATGAGCCTATGTTAGATTAGCTAGTTGGTGGGTAAGGCTCACCAAGGCTGCGATCTATAGCTGT  
TCTGAGAGGAAGATCAGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGA  
ATATTGGACAATGGGGCAACCTGATCCAGCAATACCATGTTGTGTGAAGAAGGCCTTAGGGTTGTAAAG  
CACTTTAGTCTGAGGAAGGCTATTAGGTTAAGACTAGATAGTTTACGCTTAGGCGAAGAATAAGCAC  
CGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGAGGTTGCGAGCGTTAATCGGAATTAAGTGGCGTAA  
AGGGTTCGTAGGTGGATAGATAAGTCAGATGTGAAATCCCTGGGCTCAACCTAGGAATTGCATTTGATAC  
TGTTTTATCTAGAGTTCCTAGAGGATTGGGGAATTTCCGGTGTAGCGGTGAAATGCGTAGAGATCGGAAG  
GAACATCAATGGCGAAGGCAACAATCTGGGGTTGAAGTACACTGAGGGACGAAAGCGTGGGTAGCAAAC  
AGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGAGTACTAGCTGTTGGAGTCCGTTGAAAGGCTT  
TAGTGGCGCAGCTAACCGGTTAAGTACTCCGCTGGGGACTACGGCCGCAAGGCTAAAACCTCAAAGGAAT  
TGACGGGGACCCGCAACAAGCGGTGGAGCATGTGGTTAATTCGATGCAACGCGAAGAACCTTACCTACCC  
TTGACATCCTAAGAAGAACCGAGAGATTGGTTTGTGCCCTTCGGGAACTTAGAGACAGGTGCTGCATGGCT  
GTCGTCAGTCTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTATTTTTAGTTGCCA  
TCATTAAGTTGGGCACTCTAGAGAGACTGCCGCTGATAAGGCGGAGGAAGGTGGGGACGACGCTCAAGTCA  
TCATGGCCCTTACGGGTAGGGCTACACACGTGCTACAATGGGCACTACAGAGGGCTGCCAAGGAGCGATC  
TGGAGCGAAACTCACAAAGGTGTTCTAAGTCCGGATTGATCTCTGCAACTCGAGATCATGAAGTCGGAA  
CGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCCGGGTCTTGTACACACCGCCCGTCACAC  
CATGGGAGTGGGTTGCAAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACCTTTGTGATTCATGA  
CTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCGCTG

>Gluconobacter\_cerinus\_NR041048

AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGGATCTTTTCGGGATCAGTGGCGGACGGGTG  
AGTAACCGGTAGGGATCTATCCATGGGTGGGGACAACCTCCGGGAAACTGGAGCTAATACCGCATGATAC  
CTGAGGGTCAAAGGCGCAAGTCGCCTGTGGAGGAACCTGCGTTCGATTAGCTAGTTGGTGGGGTAAAGGC  
CTACCAAGGCGATGATCGATAGCTGGTTTGGAGGATGATCAGCCACACTGGGACTGAGACACGGCCAG  
ACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGT  
GTGAAGAAGGTCTTCGATTGTAAAGCACTTTTCGACGGGGACGATGATGACGGTACCCGTAGAAGAAGCC  
CCGGCTAACTTCGTGCCAGCAGCCGCGTAATACGAAGGGGGCTAGCGTTGCTCGGAATGACTGGGCGTA  
AAGGGCGCGTAGGCGGTTTATGCAGTCAGATGTGAAATCCCCGGGCTTAACCTGGGAACTGCATTTGAGA  
CGCATAGACTAGAGGTCGAGAGAGGGTTGTGGAATTCCAGTGTAGAGGTGAAATTCGTAGATATTGGGA



AGAACACCGGTGGCGAAGGCGGCAACCTGGCTCGATACTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAA  
 CAGGATTAGATACCTGGTAGTCCACGCTGTAAACGATGTGTGCTGGATGTTGGGTAACCTTAGTTACTCA  
 GTGTGCAAGCTAACGCGCTAAGCACACCCGCTGGGGAGTACGGCCGCAAGGTTGAAACTCAAAGGAATTG  
 ACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGCAGAACCCTTACCAGGACTT  
 GCATGGGGAGGACGTACTCAGAGATGGGTATTTCTTCGGACCTCCCGCACAGGTGCTGCATGGCTGTGCT  
 CAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTCTTTAGTTGCCAGCACT  
 TTCAGGTGGGCACCTTAGAGAGACTGCCGGTGACAAGCCGGAGGAAGGTGGGGATGACGTCAAGTCCCTCA  
 TGGCCCTTATGTCTGGGCTACACACGTGCTACAATGGCGGTGACAGTGGGAAGCTACATGGCGACATGG  
 TGCTGATCTCTAAAAGCCGTCTCAGTTCGGATTGTACTCTGCAACTCGAGTACATGAAGGTGGAAATCGCT  
 AGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCCGCCGTCACACCATG  
 GGAGTTGGTTCGACCTTAAGCCGGTGGAGCAACCAGCAAGGACGCAGCCGACCACGGACGGGTACGCGACT  
 GGGGTGAAG

>Hippea\_maritima\_NR044940

GAACGCTGGCGGCGTGCCTAACACATGCAAGTCGTGGGAGAAAAGTCTCCTTCGGGAGACGAGTAAACCGG  
 CGCACGGGTGAGTAAACACGTTGGGTAACCTGCCCTGAAGTCCGGGATAACCCACCAGAAAGGTGGGCTAATA  
 CCGCATAGTTCCTTGCACCTTATCTTAGAATACAAGAGGATGGTAGAAGAGGTTATGTGTTTTGTCTTT  
 ACCTCTTTTACTCTTTTAATTCTCTGTGTTTTAAGATAAGTGCAGGGGGGAAAGGTGGCCTCTGCTTGCA  
 AGCTATCGCTTCAGGATGGGCCCGCGGCTATCAGGTAGTTGGTGGGGTAACGGCTACCAAGCTACGA  
 CGGGTAGCTGGTCTGAGAGGATGGTCAGCCACACTGGAACCTGAGACACGGTCCAGACTCCTACGGGAGGC  
 AGCAGTGGGGAATATGGGCAATGGGGGAAACCCCTGACCAGCGACGCCGCTGGAGGATGAAGGCCCTC  
 GGGTCGTAAACTCCTGTGTCAGAGGGGAAGAAGGTGCGAGGGCTAATACCCCTTTGCACCTGACGGTACCCT  
 CAGAGGAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGGCAGCGTTGTTCGGAA  
 CACTGGGCGTAAAGCGTGGTGGGCGGCTGTAAGTCCAACGTGAAAGCCTTGGGCTCAACCCAAGAAC  
 TCGTGGTGGTACTGCAGGTCTTAGAGTGGCGGAGAGGTGAGCGGAATTCGGGTGAGGGGTGAAATCCG  
 TAGATATCGGGAAGAACACCGGTGGCGAAGGCGGCTCACTGGAACGCAACTGACGCTGAGCACGAAAAGCG  
 TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGGACGCTGGGTGTCGGGAGGT  
 ACTCTTCCCGGTGCCGTAAGCTAACCGCTTAAGCGTCCCGCTGGGGAGTACGGCCGCAAGGCTAAAACCT  
 CAAAGGAATTGACGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAATTCGATGATACCGGAAGAACCT  
 TACCTGGGCTTGACATGCTGGTGGTAGTGAACCGAAAGGGGAACGACCCTTACCTTCGGGTGAGGGAGCC  
 AGCACAGGTGCTGCATGGCTGTGCTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCA  
 ACCCTCGCCCTTAGTTGCTAACGGTTTCGGCCGAGCACTCTAAGGGGACCGCCAGGGATAACCTGGAGGAA  
 GGTGGGGATGACGTCAAGTCATCACGGCCCTTATGCCAGGGCTACACACGTGCTACAATGGGGTGGACA  
 AAGGTTGCAAAACCCGCGAGGGTGGTGTAGTAAATCCCAAAACCATCCCCAGTTCGGATTGCACTCTGCAAC  
 TCGACTGCATGAAGCCGGAATCGCTAGTAAATCGCAGGTGAGTACACTGCGGTGAATACGTTCCCGGGCC  
 TTGTACACACCCCGTCCACACCATGGGAGTCCGATTCTACCGGAAGACGGTGGGCTAACCCCTTTTGGGG  
 AGGCAGCCGTCATGGTAGGGTGGCGACTGGGGTGAAGTTCGTAACAAGGTAGCCGTAGGAGAACCT

>Hymenobacter\_chitinivorans\_NR044945

AGAGTTTGGATNNTGGCTCAGGATGAACGCTAGCGGCAGGCCTAATACATGCAAGTCAACGACGAGGTTAGC  
 AATACCTTGAGTGGCGCACGGGTGCGTAACGCGTAACCAACCTACCTACATCTGGGGGATAGCCCGCGCA  
 AAGGCGGATTAATACCCGATAACCCAACAGTGTGGCATCACACAATTGGTAAAGATTTATTGGATGTAGA  
 TGGGGTTGCGTGCATTAGCTAGTTGGCGGGTAACGGCCACCAAGGCGACGATGGCTAGGGGACCTGA  
 GAGGGTGTATCCCCACACTGGCACTGAGATACGGGCCAGACTCCTACGGGAGGCGAGTAGGGAATATT  
 GGGCAATGGGCGAGAGCCTGACCCAGCCATGCCGCGTGGCGGATGAAGGCCTTCTGGGTTGTAACCGGCT  
 TTTCTCAGGGAAGAAAAAGGGGATGCGTCCCAAACCTGACGGTACCTGAGGAATAAGCACCGGCTAACTCC  
 GTGCCAGCAGCCGCGGTAATACGGAGGGTGAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGGCTAG  
 GTGGCCCGTTAAGTCCGGGGTGAAGGCCACTGCTCAACAGTGAAGTGCCTGGATACTGACGGGCTTG  
 AGTCCAGACGAGGTTGGCGGAATGGATGGTGTAGCGGTGAAATGCATAGATACCATCCAGAACCCCGATT  
 GCGTAGGCAGCTGACTAGGCTGGTACTGACACTGAGGCACGAAAGCGTGGGGAGCGAACAGGATTAGATA  
 CCCTGGTAGTCCACGCCGTAACGATGGATACTCGTTGCTAGCGATACACAGTTAGAGACTTAGCGAAAAG  
 TGTAAGTATCCCACCTGGGGAGTACGCTCGCAAGAGTGAAGTCAAAGGAATTGACGGGGGCCCGCACAA  
 GTGGTGGAGATGTGGTTTAATTCGATGATACCGGAGGAACCTTACCTAGGCTAGAATGCGCGTGAACCG  
 CTCAGAGATGAGCCTTTCTTTCGGGACACAAAGCAAGGTGCTGCTACATGGCCGCTGCTCAGTCTGCTGGA  
 GGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTACTGTTAGTTGCCAGCGGATTATGCCGGGACTC  
 TAATAGGACTGCCTGCGCAAGCAGTGGGAAGGCGGGGACGAGTCAAGTGCATGATGGCCCTTACGCCTA  
 GGGCTACACACGTGCTACAATGGACGGTACAGAGGGTTCGCTACACAGTGTGATGCCAATCTCACAAA  
 GCCGTTCTCAGTTCGGATCGGAGTCTGCAACTCGACTCCGTGAAGCTGGAATCACTAGTAATCGCGTATC  
 AGCAATGACGCGGTGAATACGTTCCCGGGCCTTGTACACACCCCGCTCAAGCCATGGAAGTTTGGTAGA  
 CCTGAAGCTGGTGTGCTCACAGAAGCCAGTTAGGGTGAACAAGTAAGTGGGGCTAAGTTCGTAACAAGG  
 TAGCCGTACCGGAAGGTGGCGGCTGGATCACCTCCTTT

>Hyphomicrobium\_methylovorum\_NR026430

GAACGAACGCTGGCGCAGGCCTAACACATGCAAGTCGAACGCCCGCAAGGGGAGTGGCAGACGGGTGA  
 GTAACGCGTGGGAACCTTCCCTATGGTACGGAATAACTCGGGGAAACTTGGAGTAATACCGTATACGCC  
 GAGAGGGGAAAGAAATTCGCTATAGGATGGCCCCGCGTCGGATTAGCTAGTTGGTGAGGTAATGGCTCAC  
 CAAGGCAGCATCCTTAGCTGGTTTGGAGAGAACGCCAGCCACACTGGGACTGAGACACGGCCAGACTC  
 CTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCGCGTGAGTGA  
 TGAAGGCCCTTAGGGTTGTAAAGCTCTTTTTGCCGGGGACGATAATGACGGTACCCGGAGAATAAGTCCCGG  
 CTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGACTAGCGTTGTTCCGAATCACTGGGCGTAAAGC  
 GCACGTAGGCGGATTTATAAGTCAGGGGTGAAATCCCGGGCTCAACCTCGGAACCTGCCTTTGATACTGT  
 GAATCTTGAGTCCGATAGAGGTGGGTGGAATTCCTAGTGTAGAGGTGAAATTCGTAGATATTAGGAAGAA  
 CACCGGTGGCGAAGCGGCCCACTGGATCGGTACTGACCTGAAAGTGCAGAAAGCGTGGGGAGCAAACAGG  
 ATTAGATACCCCTGGTAGTCCACGCCGTAACGATGGATGCTAGCCGTCCGATAGCTTGCTATTCCGGTGGC  
 GCAGCTAACGCATTAAGCATCCCGCCTGGGGAGTACGGCCGAAGGTTAAACTCAAAGGAATTGACGGG  
 GGCCCGCACAAAGCGGTGGAGCATGTGGTTAATTCGACGCAACGCGAAGAACCTTACCAGCTCTTGACAT  
 TCACGTATCGCCTGGAGAGATCCGGGAATTCAGCAATGGACACTGGGACAGGTGCATGGCTGTCTGT  
 CAGCTCGTGTCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTCGCCATTAGTTGCCATCATT  
 TAGTTGGGCACCTAATGGGACTGCCGGTGATAAGCCGGAAGAAGGTGGGGATGACGTCAAGTCATCATG  
 GCCCTTACGGGCTGGGCTACACACGTGCTACAATGGCGGTGACAATGGCAGCCACCTAGTAATAGGGAG  
 CTAATCGCAAAAAGCGTCTCAGTTCAGATTGAGGTCTGCAACTCGACCTCATGAAGTCGGAATCGCTAG  
 TAATCGCGCATCAGCATGGCGCGGTGAATACGTTCCCGGGCCTTCTACACACCGCCCGTCACACCATGGG  
 AGTTGGTCTTACCCTAAAACGGTGCCTAACCAGGAGGAGCCGGCCACGGTAAGGTGACGGACTGG  
 GGTGAAGTCGTAACAAGGTA

>Hyphomonas\_jannaschiana\_M83806

AACTTGAGAGTTTGATTCTGGCTCAGAACGAACGCTGGCGGYAGGCCTAACACATGCAAGTCGAACGAYA  
 TAGTGGNNGACGGGTGAGTAACGCGTGGGAACGTACCTTTCGCTACGGAATAGCTCTTGGAACGAGTGG  
 TAATACCGTATACGCNNTTCGGGGGAAAGATTTATCGGCGAAAGATCGGCCCGCGTTAGATTAGGTAGTT  
 GGTGGGGTAATGGCCTACCAAGCCNNGATCTATAGCTNGTCTGAGAGGATGATCAGCCACACTGGGACT  
 GAGACACGGCCNNGACTCCTACGGGAGGCAGCAGTGGGGAATCTTGACAATGGGCGAAAGCCTNATGCA  
 GCCATGCCGCGTGAATGATGAAGGCNTAGGGTTGTAAAATTCCTTTCGCCAGGGATGATAATGACAGTAC  
 CTGGNNAAGAAGCCCCGGCTNACTTCGTGCCAGCAGCNGCGGTAATACGAAGGGGGCNAGCGTTGTTTCGG  
 AATTACTGGGCGTAAAGCGCACGTAGGCGGACTTTTAAAGTCAGATGTGAAATCCCGGGGCTCAACCTCGG  
 NACTGCATTTGAAACTGGAAGTCTGGAGTTCAGGAGAGGTTAGCGGAATACCGAGTGTAGAGGTNAAAAT  
 CGTAGATATTCGGTGAACACCCAGTGGCGAAGGCGCNAACTGGACTGATACTGACGTGAGGTGCNNAA  
 GTGTGGGGAGCAAACAGGATTAGATACCCTNGTAGTCCACACCGTAAACGATGACAGCTNGTGTGTTGGCA  
 GGCATGCCNGTCCGTGACGCASSTAACGCATTAAGCTGTCCGCTGGGGAGTACGGCCGAAGGTTAAAA  
 CTCAAAGAAATTGACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAAATTCGAANNAACCGCGAGAAC  
 CTTACCTACCCTTGACATCCCGATCGCGGNTCCAGAGATGGATTCCCTCAGTTAGGCTNGATCGGNGAC  
 AGGTGCTGCATGGCNGTCGNCAGCTCGTGTGCTGAGATGTNNGGTTAAGTCCCGCAACGAGCGCAACCCT  
 CATCCTTAGTTGCCATCACGTTTGGGTGGGCNCTCTAAGGAAACTGCCGGTGGCAAGCCGGAGGAAGGTG  
 GGNATGACGTCAAGTCCCTCATNNCCTTACGGGTAGGGCTACACACGTGCTACAATGGCAGTGACAATGG  
 GATAATCCCAAAAAGCTGTCTCNGTTCAGATTGTCTCTGCAACTCGAGGGCATGAAGGTGGAATCGCTA  
 GTAATCGTGGATCAGCATGCCACGGTNAATACGTTCCCNNNNNNGTACACACNGNCCGNNACATCATGG  
 GAATNNGCTCTACNNGAAGACGCTNTGCTNACTTNGGAGG

>Ignicoccus\_islandicus\_NR044910

CGGACCCGACCGCTATCGGGGTAAGGCTAAGCCATGGGAGTCGAACGCCCGCCGCGGGCGTGGCGGA  
 CGGCTGAGTAACACGTGGCTAACCTACCCTCGGGAGGGGATAACACCGGGAAACTGGTGTAAATCCCCC  
 ATAGGGGCGGTGGCCTGGAAGGGTACCGCCCGAAAGGGGAGGCGGGGGTTATCGCCGCTCTCCGCC  
 GAGGATGCGGCTGCGCCCTATCAGGTAGTTGGCGGGGTAACGGCCCCGCAAGCCTAAGACGGGTAGGGGC  
 CGTGGGAGCGGGAGCCCCAGATGGGCACTGAGACAAGGGCCAGGCCCTACGGGGCGCACCAGGCGCGA  
 AAATCGCCCAATGCGGGCAACCGTGACGGGGTTACCCGAGTGCCTTACGGGGCTTTTCCCGCT  
 GTAAACAGGCGGGGTAATAAGCGGGGGCAAGTGTGGTGTACGCGCGGTAATACCAGCCCCGCGA  
 GTGGTGGGACGTTTATTGGGCTAAAGCGCCCGTACCAGGCGCTGGTAGGTCCCTTAAACCCGGGG  
 CTCAACCCCGGGGTGGAGGGGAAACCACAGGCTAGGGGGCGGGAGAGGCGGAGGGTACTCCCGGGGTA  
 GGGGCGAAATCCGATAATCCCGGAGGACCGCCAGTGGCGAAGGCGCTCGGCTGGAACGCGCCGACGGT  
 GAGGGGCGAAAGCCGGGGAGCGAACCGGATTAGATACCGGGTAGTCCCGGCTGTAAACGATGCGGGCTA  
 GGTGTTGGGTGGCTTCGAGCCCGCCAGTCCCGCAGGGAAGCCGTTAAGCCCGCCGCTGGGGAGTACG  
 GCCGCAAGGCTGAAACTTAAAGGAATTGGCGGGGAGCACCACAAGGGGTGGCAGGTGCGGCTTAATTGG  
 AGTCAACGCCGGGAACCTCACCGGGGGCAGCAGGATGAAGGTGAGGCTGAAGACCTTACCTGACCGG

CTGAGAGGAGGTGCATGGCCGTCGCCAGCTCGTGCCGTGAGGTGTCCGGTTAAGTCCGGCAACGAGCGAG  
 ACCCCCACCCCTACTTGCTACCCGGGGCTCCGGCCCCGGGGCACAGTAGGGGGACTGCCGCGTATAAAGG  
 CGGAGGAAGGAGGGGGCTATGGCAGGTGACATGCCCGAAACCCCGGGCTGCACGCGGGCTACAATGG  
 CGGGGACAGCGGGTTGCGACCCCGAAAGGGGGAGCAAATCCCTCAAACCCCGCCGAGGTTGGGATCGAGG  
 GCTGCAACTCGCCCTCGTGAACGCGGAATCCCTAGTAACCGCACGTTAGCATCGTGCGGTGAACACGTCC  
 CTGCTCCTTGACACACCCGCCCCGTCGCTCCACCCGAGGGGACGGGTNNNNAGGCCCCCGTAGGGAAACTC  
 CCGGGTATCCTCGAACTCCCTCCTCTCGAGGGGGGAGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTG

>*Klebsiella variicola*\_JN644499

AGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAGCACAG  
 AGAGCTTGCTCTCGGGTGACGAGCGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGA  
 TAACTACTGGAAACGGTAGCTAATACCGCATAATGTCCGAAGACCAAAGTGGGGGACCTTCGGCCTCAT  
 GCCATCAGATGTGCCAGATGGGATTAGCTGGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAG  
 CTGGTCTGAGAGGATGACCAGCCACACTGGAAGTGAACACGGTCCAGACTCCTACGGGAGGCAGCAGTG  
 GGAATATTGCACAATGGGCGCAAGCCTGATGACGGCATGCCGCGTGTGTGAAGAAGCCCTTCGGGTGT  
 AAAGCACTTTCAGCGGGGAGGAAGGCGGTGAGGTTAATAACCTCATCGATTGACGTTACCCGAGAAGAA  
 GCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGAAGCGTTAATCGGAATTACTGGGC  
 GTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCGGGCTCAACCTGGGAACTGCATTG  
 AAAGTGGCAGGCTAGAGTCTTGTAGAGGGGGTAGAATTCAGGTGTAGCGGTGAAATGCGTAGAGATCT  
 GGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGAGC  
 AAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGTGATTTGGAGGTTGTGCCCTGAGGCG  
 TGGCTTCCGGAGCTAACCGGTTAAATCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAACTCAAATGAA  
 TTGACGGGGGCCCCGACAAGCGGTGGAGCATGTGGTTAATTCGATGCAACGCGAAGAACCCTACCTGGT  
 CTTGACATCCACAGAACTTTCAGAGATGGATTGGTGCCCTTCGGAACTGTGAGACAGGTGCTGCATGGC  
 TGTCTGACGTCGTGTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTATCCTTTGTTGCC  
 AGCGGTCCGGCCGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCGAAGT  
 CATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCATATACAAAGAGAAGCGACCTCGCGA  
 GAGCAAGCGGACCTCATAAAGTATGTGCTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGGAGTCGGA  
 ATCGCTAGTAATCGTAGATCAGAATGCTACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTAC  
 ACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCCTTTGTGATTC  
 ATGACTGGGGTGAAGTCGTAACAAGGTAACC

>*Leptospira interrogans*\_DQ840043

ATGAGCACTCACTTCTCATTAAAAAGTGCATCAGTTATAACAGATTATTTATTTAAATTTTCAATTTTTT  
 CTCTTCCAGCCATTTGTTGGATTTGTTCAACTTTAATAGGTTTTGGTACTGTAAATGGACGCCTTTCTTT  
 ATTTGTTATCGGATTGTCTTTTATAATCAGTATTTTTTTATTAAAAAATATAAAGTGGAAATCTCTTCA  
 ACTTTTTCTTTTTATTAGTTATTTCTTTTTTATTAGCTTATTTCTTTTTCTATAAACTCCAAAATATGC  
 CTCAACATTTGGATGGTAAGTTAAATCCTATACTTTACGTGTTAAGGCGTTTTCTACTTTATTTTCATT  
 TTTTATTATTTTTGCACCTCCAAGTTTAAACAAAAAAGCTTTTTTTATAGGAATTGCTTTGGGAATG  
 TTTGTATTTGCAATTATCAATTCCATTGCAACCTTAGTTTATTTAGAACCCCTTATTATGAAAAGCGT  
 ATCACTTCTTTTTATAAATGGAATATAATTCGCCTGGAAATACCATTTTGGCTAGTATGCTACCTATCGT  
 TCTTTTTTGTTTTAAACGGTTATCTTTTAAAAATAGATAAAAAACTAAACTGGCAAAATGTATTTTTATA  
 TTGGTTTTCCGGATTAGCATTTCCATTTTATTTTATTAGTGCACGAACCCTTTTTTTCTGATCATTG  
 CAAATTTATTATTGGATTTCTGATCTTGTCTGTCTCGTGTCTCTCTATTTATTTTTTTCTTAAAGAAACA  
 TACATTGGTCAGAGAACTATGAACGGAATTTATCCGAAAAATTAATCATCATGTTGATTATTGGAACA  
 CGATTAATAAAGATTTTTTTATATACCCTAAAATTACAATGGATCTGAATATACTTTTTGGTATCATAA  
 TATTTTTTTGATTCGCATAAAACTTCCGGTCTATCACCGCTTTGATATTGGATAATTATTCGGTTTTT  
 ATTTTTTTAATTGCATTAATAAATCTTTAAAAGAGATTATAGATCGTTCCGATATTTTCATTTCTATA  
 TTTGTTTTATTTCCCTATTTAATGACAACAATACCTTGGGAATCCTCAGAGTCTCAAATGGTAGCATTGTT  
 TGCGGGTTTAGGGCTTTGATCACAACCTGTAGATGATCAAACCTCTGAAATGTAG

>*Lishizhenia caseinilytica*\_AB176674

AGAGTTTGATCCCGGCTCAGGATGAACGCTAGCGGCAGGCCTAGGCCTACACATGCAAGTCGAGGGGCGAG  
 CGGAGAAAAGCTTGCTTTTCTGCCGGCGACCGGCGAACGGGTGCGTAACGCGTATACAATCTGCCCTTGTA  
 CAGGAGGATAGCCCGGAGAAAATTCGGATTAATACTCCATAGCATTATCGAATCGCATGATTGATTCTTA  
 AAATTCGGTGGTACAAGATGAGTATGCGTCCTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAA  
 CGATAGGTAGGGGCTGAGAGGATTATCCCCACACTGGTACTGAGACACGGACCAGACTCCTACGGGA  
 GGCAGCAGTGAGGAATATTGGTCAATGGACGAAAGTCTGAACCAGCCATGCCGCGTGCAGGAAGAATGCC  
 CTATGGGTTGTAACCTGCTTTTTATTTGGGAAGAACTTCTTACGTGTAGGAAGCTGACGGTACCAAACG  
 AATAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGAAGCGTTATCCGGAATCAT

TGGGTTTAAAGGGTCCGCAGGCGGGCGTATAAGTCAGTGGTGAAATCTCTCGGCTCAACCGAGAACTGC  
 CATTGATACTGTATGCTTGAATTCGGTCAAGTAGGCGGAATGAGTAGTGTAGCGGTGAAATGCATAGA  
 TATTACTCAGAACACCGATAGCGAAGGCAGCTTACTAGGCCTGGATTGACGCTGAGGGACGAAAGCGTGG  
 GGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGATTACTCGATATCAGCGATATAC  
 TGTTGGTGTCTAAGCGAAAAGTGATAAGTAATCCACCTGGGGAGTACGATCGCAAGGTTGAAACTCAAAGG  
 AATTGACGGGGGCCGACAAGCGGTGGAGCATGTGGTTTAAATTCGATGATACGCGAGGAACCTTACCAG  
 GGCTTAAATGCAGAACGACCGGTCTGGAAACAGACCTTCTTCGGACGGTTTGCAAGGTGCTGCATGGCT  
 GTCGTCAGCTCGTGCCGTGAGGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTATCTTTAGTTGCTA  
 GCAGGTAATGCTGAGGACTCTAAAGAACTGCCAGCGCAAGCTGAGAGGAAGCGGGGACGACGTCAAGT  
 CATCACGGCCCTTACGTCTGGGCCACACACGTGCTACAATGGTCAGTACAGAGGGCAGCTACCTAGCGA  
 TAGGATGCGAATCTCGAAAGCTGATCTCAGTTCGGATTGGAGTCTGCAACTCGACTCTATGAAGCTGGAA  
 TCGCTAGTAATCGCGTATCAGCCATGACGCGGTGAATACGTTCCCGGGCCTTGTACACACCCCGCGTCAA  
 GCCATGGAAGCTGGGGTGCCTGAAGTCGGTAACCGCAAGGAGCTGCCTAGGGTAAAAC TAGTAACTGGG  
 GCTAAGTCGTAACAAGGTAGCCGTACCGGAAGGT

>Lysobacter\_koreensis\_NR041014

CCCCCTCGAGTGAACGCTGGCGGCAGGCCCTAACACATGCAAGTCGAACGGCAGCACAGTAAGAGCTTGCT  
 CTTATGGGTGGCGAGTGGCGGACGGGTGAGGAATACGTCGGAATCTGCCTATTTGTGGGGGATAACGTAG  
 GGAAACTTACGCTAATACCGCATAACGACCTACGGGTGAAAGCGGAGGACCTTCGGGCTTCGCGCAGATAG  
 ATGAGCCGACGTCGATTAGCTAGTTGGCGGGGTAAAGGCCACCAAGGCGACGATCCGTAGCTGGTCTG  
 AGAGGATGATCAGCCACACTGGAAGTGAAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGAATAT  
 TGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAGCACT  
 TTTGTCCGAAAGAAAAGCACTCGGTTAATACCCGGGTGTCATGACGGTACCGGAAGAATAAGCACCGGC  
 TAACCTCGTGCCAGCAGCCGCGGTAATACGAAGGTTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCG  
 TGCGTAGTGGTTTGTAAAGTCTGATGTGAAAGCCCTGGGCTCAACCTGGGAATGGCATTGGATACTGGC  
 AATCTAGAGTGCCTGAGAGGGGTGTTGGAATTCCTGGTGTAGCAGTGAATGCGTAGATATCGGGAGGAAC  
 ATCTGTGGCGAAGGCGACACCCTGGACCAGCACTGACACTGAGGCACGAAAGCGTGGGGAGCAAAACAGGA  
 TTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGATGTTGGGAGCAACTAGGCTCTCAGTAT  
 CGAAGCTAACCGGTTAAGTTCGCCGCTGGGAAGTACGGTTCGCAAGACTGAAACTCAAAGGAATTGACGG  
 GGGCCCGCACAAGCGGTGGAGTATGTGGTTTAAATTCGATGCAACGCGAAGAACCTTACCTGGCCTTGACA  
 TGTGCGAATCCTTGAGAGATCGAGGAGTGCCTTCGGGAACGCGAACACAGGTGCTGCATGGCTGTCGTC  
 AGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTTGTCTTAGTTGCCAGCACGT  
 AATGGTGGAACTCTAAGGAGACCGCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAGTCATCAT  
 GGCCCTTACGGCCAGGGCTACACACGTAATAAGTGGTAGGGACAGAGGGTCGCAAACTCGCGAGAGCCA  
 GCCAATCCAGAAACCCCTATCTCAGTCCGGATCGGAGTCTGCAACTCGACTCCGTGAAGTCGGAATCGCT  
 AGTAATCGCAGATCAGCATTGCTGCGGTGAATACGTTCCCGGCCTTGTACACACCCCGCCCTCACACCAT  
 GGGAGTTTGTGACCAGAAGCAGGTAGCTTAACTTCGGGAGGGCGCTTGCCACGGTGTGCCAGATGAC  
 TGGG

>Macrococcus\_bovicus\_NR044928

AGAGTTTGTATNNGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGGACAGACGA  
 GGTGCTTGACCTCTGAAGTCAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTACCTGTAAGACTGGG  
 ATAACCTCGGAAACCGGAGCTAATACCGGATAATATTTCCACCTCATGGTGAATAGTGAAAGACGGT  
 TTTGCTGTCACTTACAGATGGACCCGCGGCGCATTAGCTAGTTGGTGAAGTAACGGCTACCAAGGCGAC  
 GATGCGTAGCCGACCTGAGAGGGTATCGGCCACACTGGGACTGAGACACGCGCCAGACTCCTACGGGAG  
 GCAGCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAAGTGAAGAAGGTCT  
 TCGGATCGTAAAACCTGTTGTAAGGGAAGAACAAGTACGTTAGTAACTGAACGTACCTTGACGGTACCT  
 TACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGTAATACGTAGGTGGCAAGCGTTATCCGGAA  
 TTATTGGGCGTAAAGCGCGGTAGGCGGTTTCTTAAAGTCTGATGTGAAAGCCCCCGCTCAACCGGGGAG  
 GGTCAATTGGAACCTGGGAGACTTGAGTGCAGAAGAGGAGAGTGAATTCATGTGTAGCGGTGAAATGCG  
 CAGAGATATGGAGGAACACCAGTGGCGAAGGCGGCTCTCTGGTCTGTAACGTGACGCTGAGGTGCGAAAGC  
 GTGGGGATCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGTAAAGTGTGGGGGG  
 TTTCCAGCCCTCAGTGTGACGCTAACGCATTAGCACTCCGCTGGGGAGTACGGTCGCAAGACTGAA  
 CTCAAAGGAATTGACGGGGACCCGCAACGCGGTGGAGCATGTGGTTTAAATTCGAAGCAACGCGAAGAAC  
 CTTACCAAATCTTGACATCCTCTGACAACCTGAGAGCAGAGCGTTCCCTTCGGGGGACAGAGTGCAG  
 GTGGTGCATGGTTGTCGTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTA  
 TCTTTAGTTGCCATCATTCAGTTGGGCACTCTAGAGAGACTGCCGCTGACAAACCGGAGGAAGGTGGGGA  
 TGACGTCAAATCATCATGCCCTTATGATTTGGGCTACACACGTGCTACAATGGATGGTACAAAGGGCAG  
 CAAAACCGGAGGTCAAGCAAATCCATAAAACCATCTCAGTTCGGATTGTAGTCTGCAACTCGACTAC  
 ATGAAGCTGGAATCGCTAGTAATCGTAGATCAGCATGCTACGGTGAATACGTTCCCGGGTCTGTACACA  
 CCGCCCGTACACCACGAGAGTTTGTAAACCCGAAGCGGTGGAGTAACCTTTACAGGAGCTAGCCGTC

GAAGGTGGGACAGATGATTGGGGTGAAGTCGTAACAAGGTAGCCGTATCGGAAGGTGCGGCTGGATCACC  
TCCTTT

>*Microbulbifer\_variabilis*\_NR041021

ATAGAGTTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCCTAACACATGCAAGTCGAGCGCGAAAG  
TTCTTCGGAATGAGTAGAGCGGCGGACGGGTGAGTAACGCGTGGGAAATTGCCAGTAGTGGGGGACAAC  
ATTTCGAAACGGATGCTAATACCGCATAACGCCCTACGGGGGAAAGCAGGGGATCTTCGGACCTTGTGCTA  
TTGGATATGCCCGCGTCGGATTAGCTAGTTGGTGAAGTAATGGCTCACCAAGGCAACGATCCGTAGCTGG  
TCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGA  
ATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGCTCTAGGGTTGTAAG  
CACTTTTCAGTAGGGAGGAAGGCCCTTAAAGTTAATACCTTTGAGGATTGACGTTACCTACAGAAGAAGCAC  
CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAGAGCGTTAATCGGAATTACTGGGCGTAA  
AGCGCGCGTAGGCGGTTAGTTAAGCTGGATGTGAAAGCCCCGGGCTCAACCTGGGAAGTGCATTCAGAAC  
TGGCTGGCTAGAGTACGAGAGAGGGTAGTGGAAATTTCCGTGTAGCGGTGAAATGCGTAGATATAGGAAG  
GAACATCAGTGGCGAAGGCGACTGCCTGGCTCGATACTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAAC  
AGGATTAGATACCCTGGTAGTCCACGCCGTAACGATGTCTACTAGTCGTAGGGTTCCCTGAGGACTTTG  
TGACGCAGCTAACGCAATAAGTAGACCCGCTGGGGAGTACGGCCGCAAGGTTAAACTCAAATGAATTGA  
CGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGGCTTG  
ACATCCTCGGAAGTCTGCAGAGATGCGGATGTGCCCTTCGGGAACCGAGTGACAGGTGCTGCATGGCTGTC  
GTCAGCTCGTGTCTGATGATGTTGGGTTAAGTCCCCTAACGAGCGCAACCCTTGTCTTAGTTGCCAGCA  
CGTAATGGTGGGAACCTTAGGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGACGACGTCAGTCAAT  
CATGGGCCCTTACGTCCTGGGCTACACACGTGCTACAATGGTTGGTACAGACGGTTCGCTAAGCCGCGAGGT  
GGAGCTAATCCGAAAAAACCAATCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATC  
GCTAGTAATCGTGAATCAGAATGTCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTACACC  
ATGGGAGTGGGTTGCTCCAGAAGTGGCTAGTCTAACCTTCGGGGGACGGTCACCACGGAGTGATTCATG  
ACTGGGGTGAAGTCGTAACAAG

>*Myroides\_pelagicus*\_NR041042

GAGTTTTGATCCTGGCTCAGGATGAACGCTAGCGGCAGGCCCTAACACATGCAAGTCGAGGGGTATAATTA  
GCTTGCTAATTAGAGACCGGCGCACGGGTGAGTAACGCGTATGCAACCTACCTATTACAGGGGAATAGCC  
AGAAGAAATTTCTGATTAATGCTCCATGGTTTTACTTGAATGGCATCATTTGATTAATAAAGATTTATCGGT  
AATAGATGGGCATGCGTGTCAATTAGCTAGTTGGTATGGTAACGGCATAACCAAGGCAACGATGACTAGGGG  
TCCTGAGAGGGAGGTCCCCACACTGGTACTGAGACACGGGACCAGACTCCTACGGGAGGCAGCAGTGA  
GAATATTGGTCAATGGAGGCAACTCTGAACCGCATGCCGCGTGCAGGATGACGGTCTCATGGATTGTA  
AACTGTTTTGTACAGGAAGAAATGTTACTACGTAGTAAATTTGACGGTACTGTAAGAATAAGGATCGG  
CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGAATTATTGGGTTTAAAGG  
GTTTCGTAGGCGGTTGGATAAGTCAGTGGTGAATCTCATAGCTTAACTATGAAACTGCCGTTGATACTGT  
CTGACTTGAATAGTATGGAAGTAACTAGAATATGTAGTGTAGCGGTGAAATGCTTAGATATTACATGGAA  
TACCAATTGCGAAGGCAGGTTACTACGTACTTATTGACCGCTGATGAACGAAAGCGTGGGTAGCGAACAG  
GATTAGATACCCTGGTAGTCCACGCCGTAACGATGGATACTAGCTGTTCCGGACTTCGGTTTTGAGTGGCT  
AAGCGAAAGTGATAAGTATCCACCTGGGGAGTACGTTCCGAAGAATGAAACTCAAAGGAATTGACGGGG  
GCCCGCACAAAGCGGTGGAGCATGTGGTTTAATTCGGATGATACGCGAGGAACCTTACCAGGGCTTAAATG  
TAGATTGACAGTTTTGGAAACAGACTTTTTCTTCGGACAATTTACAAGGTGCTGCATGGTTGTCGTAGCT  
CGTGCCGTGAGGTGTCAGGTTAAGTCTATAACGAGCGCAACCCCTATTGTTAGTTACCAGCGCGTTAAG  
GCGGGGACTCTAGCAAGACTGCCGGTGCAACCCGTGAGGAAGGTGGGGATGACGTCAAATCATCCACGGC  
CCTTACGTCCTGGGCTACACACGTGCTACAATGGCAAGTACAGAAAGCAGCTACCTGGCAACAGGATGC  
GAATCTCAAAGCTTGTCTCAGTTCCGATTGGAGTCTGCAACTCGACTCTATGAAGCTGGAATCGCTAGT  
AATCGGATATCAGCCATGATCCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCAAGCCATGGA  
AGCTGGGGGTACCTGAAGTCGGTGACCGCAAGGGAGCTGCCTAGGGTAAAAGTACTGACTGGGGCTAAGT  
CGTAACAAGGGTAACC

>*Myxococcus\_fulvus*\_AJ233917

GTTTTGATCCTGGCTCAGAGCGAACGCTGGCGGCGTGCCTAACACATGCAAGTCGAGCGCGAATAGGGGCA  
ACCCCTTAGTAGAGCGGCGCACGGGTGCGTAACACGTGGATAATCTGCCTGGATGCTCGGGATAACAGTCT  
GAAAGATTGGCTAATACCGGATAAGCCACGGTTTTCTTCGGAGACTGAGGGAAAAGGTGGCCTCTGTATA  
CAAGCTATCACAACCAGATGAGTCCGCGGCCATCAGCTAGTTGGCGGGGTAATGGCCACCAAGGCAAC  
GACGGGTAGCTGGTCTGAGAGGACGATCAGCCACACTGGAAGTGAAGACACGGTCCAGACTCCTACGGGAG  
GCATCAGTGGGGAATTTTGGCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGTGTGATGAAGGTCT  
TTGGATTGTAAAGCACTTTCGACCGGGAAGAAAACCCGTTGGCTAACATCCAACGGCTTGACGGTACCGG  
GAGAAGAAGCACCGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGGTGAAGCGTTGTTCGGAAT  
TATTGGGCGTAAAGCGCGTGTAGGCGGCGTGACAAGTCGGGTGTGAAAGCCCTCAGCTCAACTGAGGGAAG

TGCGCCCGAAACTGTCGTGCTTGAGTGCCGGAGAGGGTGGCGGAATTCCCCAAGTAGAGGTGAAATTCGT  
 AGATATGGGGAGGAACACCGGTGGCGAAGGCGGCCACCTGGACGGTAACCTGACGCTGAGACGCGAAAGCG  
 TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAACGATGAGAAGTAGGTGTCGTGGGAG  
 TTGACCCCCGCGGTGCCGAAGCTAACGCATTAAGTTCTCCGCCTGGGAAGTACGGTCGCAAGACTAAAAAC  
 TCAAAGGAATTGACGGGGGCCCCGACAAGCGGTGGAGCATGTGGTTAATTTCGACGCAACGCGCAGAACC  
 TTACCTGGTCTTGACATCCTCGAATGCCCTCAGAGATGAGGCGGTGCCCGCAAGGGAGCCGAGAGACAGG  
 TGCTGCATGGCTGTGTCAGCTCGTGTGTCGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTCGC  
 CTTTAGTTGCCACGCAAGTGGATCTCTAGAGGGACTGCCGGTGTAAACCAGGAGGAAGGTGGGGATGACG  
 TCAAGTCTCATGGCCTTATGACCAGGGCTACACACGTGCTACAATGGCCGGTACAGAGCGTTGCCAAC  
 CCGCGAGGGGGAGCTAATCGCATAAAACCGGTCTCAGTTCAGATTGGAGTCTGCAACTCGACTCCATGAA  
 GCGGAATCGCTAGTAATCGCAGATCAGCACGCTGTGGTGAATACGTTCCCGGGTCTTGTACACACCGCC  
 CGTCACACCATGGGAGTCGATTGCTCCAGAAGTCACTTACCAAGAGGTGCCAAGGAGTGGTCGGTAAC  
 TGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCCGGTGGATCACCTCC

>Nannocystis\_exedens\_AJ233946

TTTGATCCTGGCTCAGAGCGAACGTTTTCGGCGGGCCTAACACATGCAAGTCGAACGGGCTAGCAATAGT  
 CAGTGGCGCACGGGTGCGTAACACGTAGGTAATCAACCCTTGGTTCGGGATAACGTTCTGAAAGGAGCG  
 CTAATACCGGACGCGTCTTCGGGAGCTTCGGCTCTTGTCGAGAAAAGACCCGCAAGGGTTGCCGAGGGACG  
 AGCCTGCGGCCCATCAGCTAGTTGGCGAGGTAATAGCTCACCAAGGCGAAGACGGGTAGCTGGTCTGAGA  
 GGATGATCAGTCACACTGGAAGTGGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGC  
 GCAATGGGCGAAAGCCTGACGCAGCCACGCCGCTGAGCGATGAAGGCCTTCGGGTGCTAAAGCTCTGTG  
 GGGAGAGACGAAGAAAGCCTGTGAAGAGCAGGCCTTACGGTATCTCCTTAGCAAGCACCGGCTAACTCC  
 GTGCCAGCAGCCGCGGTAATACGGAGGGTGCACGCTTGTCTCGGAATCATTGGGCGTAAAGCGCACGTA  
 GCGCGCGGCTAAGCGGGATGTGAAAGCCCAGGGCTCAACCCTGGAAGTGCATCCCGAAGTGTGTCGCTTG  
 AATCTCGGAGGGGACAGAGAATTCCCGGTGTAGAGGTGAAATTCGTAGATATCGGGAGGAATACCAGTG  
 GCGAAGGCGCTGTCCCTGGACGAAGATTGACGCTGAGGTGCGAAAAGCGTGGGGAGCAAACAGGATTAGATA  
 CCCTGGTAGTCCACGCTGTAAACGATGAGTGTGGACGGTGGAGGATTTGACCCCTTCGCTGTGCAAGCT  
 AACCGTTAAGCACTCCGCCTGGGGAGTACGGTTCGCAAGACTAAAACCTCAAAGGAATTGACGGGGCCCG  
 CACAAGCGGTGGAGCATGTGGTTAATTTCGACGCAACGCGCAGAACCCTTACCTGGGTAAATCCACTGGA  
 ACCTGGCTGAAAGGCTGGGGTGCCTTCGGGGAGCCGGTGAAGGTGCTGCATGGCTGTGTCAGCTCG  
 TGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTATCGCCAGTTGCCACCATGAGTTGG  
 GAACTCTGGCGAGACTGCCGGTCTAAAACCGGAGGAAGGTGGGGACGACGTCAGTCCATGGCCCTTA  
 TGCCAGGGCTACACACGTGCTACAATGGCTGGTACAAAGAGCCGCAAGCCCGCGAGGGTGAAGAAATCT  
 CAAAAACAGCTCAGTTCGGATTGCAGTCTGCAACTCGACTGCATGAAGCTGGAATCGCTAGTAATCG  
 GAGATCAGCAGCTCCGGTGAATACGTTCCCGGGCTTGTACACACCGCCCGTACACCATGGGAGTCGG  
 CTGCTCCAGAAGTAGGAACCTCAACCGCAAGGAAAGGCCCTACCAAGGAGCGGTGACTGGGGTGAA  
 GTCGTAACAAGGTTGCCGTAGGGGAACCTGCCGGTGGATCACCTC

>Neisseria\_gonorrhoeae\_X07714

TGAACATAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCTTTACACATGCAAGTCGGACGG  
 CAGCACAGGGAAGCTTGCTTCTCGGGTGGCGAGTGGCGAACGGGTGAGTAACATATCGGAACGTACCGGG  
 TAGCGGGGGATAACTGATCGAAAGATCAGCTAATACCGCATACGTCTTGAGAGGGAAAGCAGGGGACCTT  
 CGGGCCTTGCCTATCCGAGCGGCCGATATCTGATTAGCTGGTTGGCGGGGTAAGGCCACCAAGGCGA  
 CGATCAGTAGCGGGTCTGAGAGGATGATCCGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGA  
 GGCAGCAGTGGGGAATTTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCTGTCTGAAGAAGGCC  
 TTCGGGTTGTAAAGGACTTTTGTGAGGGAAGAAAAGGCTGTTGCCAATATCGCGCGCCGATGACGGTACC  
 TGAAGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGGCAGCGTTAATCGGA  
 ATTACTGGGCGTAAAGCGGGCGCAGACGGTTACTTAAGCAGGATGTGAAATCCCCGGGCTCAACCGGGA  
 ACTGCGTTCGAACTGGGTGACTCGAGTGTGTCAGAGGGAGGTGGAATTCACGTGTAGCAGTGAATGC  
 GTAGAGATGTGGAGGAATACCGATGGCGAAGGCAGCCCTCGGGATAACACTGACGTTTATGTCGGAAAG  
 CGTGGGTAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTCAATTAGCTGTTGGGCA  
 ACTTGATTGCTTGGTAGCGTAGCTAACGCGTGAATTTGACCGCTGGGGAGTACGGTCGCAAGATAAAA  
 CTCAAAGGAATTGACGGGGACCCGCACAAGCGGTGGATGATGTGGATTAATTTCGATGCAACGCGAAGAAG  
 CTTACCTGGTTTTGACATGTGCGGAATCCTCCGGAGACGGAGGAGTGCCTTCGGGAGCCGTAACACAGGT  
 GCTGCATGGCTGTGTCAGCTCGTGTGTCGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTG  
 ATTAGTTGCCATCATTCGGTTGGGCACTCTAATGAGACTGCCGGTGACAAGCCGGAGGAAGGTGGGGATG  
 ACGTCAAGTCCCTCATGGCCCTTATGACCAGGGCTTCACACGTCATACAATGGTGGTACAGAGGGTAGCC  
 AAGCCGCGAGGGCGAGCCAATCTCACAAAACCGATCGTAGTCCGGATTGCACTCTGCAACTCGAGTGCAT  
 GAAGTCGGAATCGCTAGTAATCGCAGGTGAGCATACTGCCGTGAATACGTTCCCGGGTCTTGTACACACC  
 GCCCGTACACCATGGGAGTGGGGGATACAGAAGTAGGTAGGGTAACCGCAAGGAGTCCGCTTACCACG  
 GTATGCTTCATGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCCGGTGGATCACCTCCT

TTCT

>*Nocardia\_anaemiae*\_NR041010

ATCCTGGCTCAGGACGAACCGCTGGCGGCGTGCTTAACACATGCAAGTCGAGCGGTAAGGCCCTTCGGGGT  
ACACGAGCGGCGAACCGGGTGAGTAACACGTGGGTGATCTGCCCTGTACTTCGGGATAAGCCTGGGAAACT  
GGGTCTAATACCGGATATGACCACGGGATGCATGTCTTGTGGTGGAAAGATTTATCGGTGCAGGATGGGC  
CCGCGGCCATCAGCTTGTGGTGGGGTAATGGCTACCAAGGCGACGACGGGTAGCCGACCTGAGAGGG  
TGACCGGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACA  
ATGGGCGGAAGCCTGATGCAGCGACGCCGCTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCGAC  
AGGGACGAAGCGTAAGTGACGGTACCTGTAGAAGAAGCACCGGCCAACTACGTGCCAGCAGCCGCGGTAA  
TACGTAGGGTGCAGCGTGTCCGGAATTACTGGGCGTAAAGAGCTTGTAGGCGGTTTCGTGCGTCGATC  
GTGAAAACCTGGCGGCTCAACTGCCAGCTTGGCGTGCATACGGGCGGACTAGAGTACTTCAGGGGAGACTG  
GAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTCTGGG  
AAGTAACTGACGTGAGAAGCGAAAGCGTGGGTAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCGT  
AAACGGTGGGTACTAGGTGTGGGTTTCCTTCCACGGGATCCGTGCCGTAGCTAACGCATTAAGTACCCCG  
CCTGGGGAGTACGGCCGCAAGGCTAAAACCTCAAAGGAATTGACGGGGGCCCGCACAAAGCGCGGAGCATG  
TGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGTTTGACATACACCAGAAACATCCAGAGATGGG  
TGCCCCCTTGTGGTTGGTGTACAGGTGGTGCATGGCTGTCGTGAGTCTGTGTCGTGAGATGTTGGGTTAA  
GTCCCGCAACGAGCGCAACCTTATCTTATGTTGCCAGCGCTTATGGCGGGACTCGTGAGAGACTGCC  
GGGTCAACTCGGAGGAAGGTGGGGACGACGTCAAGTCATCATGCCCTTATGTCCAGGGCTTCACACAT  
GCTACAATGGCCGTTACAGAGGGCTGCGATACCGTGAGGTGGAGCGAATCCCTTAAAGCCGGTCTCAGTT  
CGGATCGGGGTCTGCAACTCGACCCCGTGAAGTTGGAGTCGCTAGTAATCGCAGATCAGCAACGCTGCGG  
TGAATACGTTCCCGGCCCTTGTACACACCGCCCGTACGTCATGAAAGTCGGTAACACCCGAAGCCGGTG  
GCCTAACCCCTTGTGGGAGGGAGCCGTCGAAGGTGGGATTGGCGATTGGGACGAAGTCGTAACAAGGTAG  
CCGTACCGGAAGGTGCGGCTGGATCACT

>*Novosphingobium\_naphthalenivorans*\_AB684349

AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCGTGCCCTAACACATGCAAGTCGAACGAGACCTTCG  
GGTCTAGTGGCGCACGGGTGCGTAACGCGTGGGAATCTGCCCTTGGGTTCCGGAATAACAGTGAGAAATTA  
CTGCTAATACCGGATGATGTCTTCGGACCAAAGATTTATGCCCCAAGGATGAGCCCGCGTAGGATTAGCT  
AGTTGGTGGGGTAAATGGCCTACCAAGGCGACGATCCTTAGCTGGTCTGAGAGGATGATCAGCCACACTGG  
GACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGA  
TCCAGCAATGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTTACCAGGGATGATAATGACA  
GTACCTGGAGAATAAGCTCCGGCTAACTCCGTGCCAGCAGCCCGGTAATACGGAGGGAGCTAGCGTTGT  
TCGGAATTACTGGGCGTAAAGCGCGCTAGGCGTTACTCAAGTCAGAGGTGAAAGCCGGGGCTCAACC  
CCGGAACTGCCTTTGAAACTAGGTGACTAGAATCTTGGAGAGGTGAGTGAATTCAGATTGTTCTGCAACTC  
AATTCGTAGATATTCGGAAGAACCAGTGGCGAAGGCGACTGACTGGACAAGTATTGACGCTGAGGTGC  
GAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAACGATGATAACTAGCTGTC  
CGGGCACATGGTGTTTGGGTGGCGCAGCTAACGCATTAAGTTATCCGCCTGGGGAGTACGGTCGCAAGAT  
TAAAACCTCAAAGGAATTGACGGGGCCCTGCACAAGCGGTGGAGCATGTGGTTTAAATTCGAAGCAACGCGC  
AGAACCTTACCAGCGTTTACATCCTGATCGCGAATAGCAGAGATGCTTTTCTTCAGTTCCGCTGGATCA  
GTGACAGGTGCTGCATGGCTGTGCTGAGTCTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCA  
ACCCTCGTCTTAGTTGCCATCATTTAGTTGGGCACCTAAGGAAACTGCCGGTGATAAGCCGGAGGAAG  
GTGGGGATGACGTCAAGTCCCTCATGGCCCTTACACGCTGGGCTACACACGTGCTACAATGGCGGTGACAG  
TGGGCAGCGAGCACGCGAGTGTGAGCTAATCTCCAAAAGCCGTCTCAGTTCCGATTGTTCTGCAACTC  
GAGAGCATGAAGGCGGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCAGGCTTG  
TACACACCGCCCGTACACCATGGGAGTTGGGTTACCCGAAGGCGTTGCGCTAACTCGCAAGAGAGGCA  
GGCGACCAGGTGGGCTTAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCGGCTGG  
ATCACCTCCTT

>*Pelotomaculum\_schinkii*\_NR044877

TGGATAACCTGCCTTAATGACCGGGATAACGCCGGGAAACTGGCGCTAATACCGGATACGCTCACGGAAA  
ACACATGTTTGGGTAAGGAAAGGAGCAATCCGCTTAAGATGGGTCGCGTCCCATTAGCTAGTTGGAGG  
TGTAAGAGTACCCCAAGGCGAGATGGGTAGCCGGCTGAGAGGGTGGACGGCCACACTGGAACCTGAGAG  
ACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAAC  
GCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAAATCTGTCTTTCAGGGAAGAAGAAAGTGACGGTACCT  
GAGGAGGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGTAAAACGTAGGGGGCGAGCGTTGTCCGGAA  
TTACTGGGCGTAAAGGGCGGTAGGCGGTTTACTAAGTCTTATGGTGAAGAACTATCGGCTCAACCGGTAG  
CGTGCCTGAGAACTGGTAGACTTGAGGGCAGGAGAGGGGAGTGAATTCAGTGTAGCGGTGAAATGC  
GTAGATATTGGGAGGAACACCAGTGGCGAAGGCGGCTCTTGGCCTGTTACTGACGCTGAGGCGCGAAAG  
CGTGGGGAGCGAACGGGATTAGATACCCCGGTAGTCCACGCCGTAACGATGGGTGCTAGGTGTAGGAGG

TATCGACCCCTTCTGTGCCGTAGTTAACACAATAAGCACCCCGCCTGGGGAGTACGGCCGCAAGGTTGAA  
 ACTCAAAGGAATTGACGGGGGCCCCGACAAAGCGGTGGAGCATGTGGTTAATTTCGACGCAACGCGAAGAA  
 CCTTACCAGGGTTTGACATCCTCTGACAGCCTATGAAAGTAGGTTTTCTATCTTCGGATGGACAGGGAGA  
 CAGGTGGTGCATGGTTGTCGTGAGCTCGTGTCTGTGAGATGTTGGGTTAAGTCCCCGCAACGAGCGCAACCC  
 CTACGTTTTAGTTGCTAACCGGTGAAGGCGAGCACTCTAGAGGAACTGCCGTTGACAAAACGGAGGAAGGT  
 GGGGATGACGTCAAATNATCATGCCCCCTTATGTCTGGGCTACACACGTGCTACAATGGCCGGTACAAAG  
 GGAAGCGAAGTCGCGAGGCGGAGCGAATCCCAAAAAGCCGGTCTCAGTTCGGATTGCAGGCTGCAATTTCG  
 CCTGCATGAAGTCGGNATCGCTAGTAATCGCAGGTGAGCATACTGCGGTGAATACGTTCCCGGGCCTTGT  
 ACACACCGCCCGTACACCACGAAAGCTGACAACACCCCGAAGCCGGTGACTTAACCTGCAAAGGAGAGN  
 GCCGTGCAAGGTGGGGTTGGTGATTGGGGTGA

>Polyangium\_cellulosum\_M94282

NTTAACTGGAGAGTTTATCCTGGCTCAGAACGAACGTTAGCGGGCGCGCTTAACACATGCAAGTCGAGCG  
 AGAAAGGGCTTCGGCCCCGGTAAAGCGGCGCACGGGTGAGTAACACGTAGGTAATCTGCCCCAGGTGGT  
 GGATAACGTTCCGAAAGGAGCGCTAATACAGCATGAGACCACGTCTTCGAAAGAGGATGAGGTCAAAGCC  
 GGCCTCTTACGAAAGCTGGCGCCAGGGGATGAGCCTGATGATCAGCCACACTGGAAGTGAAGACCGGTCNAGA  
 TACCAAGGCGAAGACGGGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGAAGTGAAGACCGGTCNAGA  
 CTCCTACGGGAGGCAGCAGTGGGGAATCTTGCGCAATGGGCGAAAGCCTGACGCAGCGACGCCGCGTGAG  
 TGATGAAGGCCCTTCGGGTTGTAAAGCTCTGTGGAGGGGACGAATAAGGGTTGGCTAACATCCAGCTCGA  
 TGACGGTACCCCTTTAGCAAGCACCGGCTAACTCTGTGCCAGNAGCCGCGGTAAGACAGAGGTTGCAAAC  
 GTTGTTCGGAATTACTGGGCGTAAAGCGCATGTAGGCGGTTTCGTAAAGTCAGATGTGAAAGCCCTGGGCT  
 TAACCCAGGAAGTGCATTTGAAACTCACGAACCTGAGTCCCGGAGAGGAAGGCGGAATTTCTCGGTGTAGA  
 GGTGAAATTCGTAGATATCGAGAGGAACATCGGTGGCGAAGGCGGCCTTCTGGACGGTGACTGACGCTGA  
 GATGCGNAAGCGTGGGGAGCAAACAGGATTAGATACCCCTGGTAGTCCACGCCGTAACGATGGGTGCTAG  
 GTGTGCGGGCTTTGACTCCTGCGGTGCCGTAGCTAACGCATTAAGCACCCCGCCTGGGGAGTACGGCCG  
 CAAGGCTAAAACCTCAAAGGAATTGACGGGGGCCNGCAAGCGGTGGAGCATGTGGTTCAAATTCGANNA  
 ACGCGCAGAACCTTACCTGGGCTAGAAAATGCAGGNACCTGGTTGAAAGATCGGGGTGCTCTTCGGAGAA  
 CCTGTAGTTAGGTGCTGCATGGCTGTCTGTCAGCTCGTGTCTGTGAGATGTTGGGTTAAGTCCCAGCAACGAG  
 CGCAACCCCTATCGTTAGTTGCCAGCGGTTYGGCCGGGCACTCTAGCGAGACTGCCGATATTTAAATCGG  
 AGGAAGGTGGGGATGACGTCAAGTCCCTCATGGCCCTTATGTCCAGGGCTACACACGTGCTACAATGGGCG  
 GTACAGACGGTTCGCGAACC CGGAGGGGAAGCCAATCCGAAAAAACCGTCTCAGTACGGATAAGAGTCT  
 GCAACTCGACTCTTTGAAGTTGGAATCGCTAGTAATCCCTGATCAGCAGGCAGGGGTGAATACGTTCCCG  
 GGCNTTGTACACACCGCCCGTACACCATGGGAGTCGATTGCTCCAGAAGTGGCTGCGCCAACCCGCAAG  
 GGAGGCAGGCCCCCAAGGAGTGGTTGGTAACTGGGGNNNNNNNGTAACAAGNNNNNNNNNNNNNNNNNN  
 NNNNNNGATCACCTCCTTCT

>Providencia\_rettgeri\_JN644501

CTGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCAACACATGCAAGTCGAGCGGTAACAGG  
 GGAAGCTTGCTTCCCGCTGACGAGCGGCGGACGGGTGAGTAATGTATGGGGATCTGCCCGATAGAGGGGG  
 ATAACTACTGGAAACGGTAGCTAATACCGCATAATCTCTCAGGAGCAAAGCAGGGGAACCTCGGTCCCTG  
 CGCTATCGGATGAACCCATATGGGATTAGCTAGTAGGTGAGGTAATGGCTCACCTAGGCGACGATCCCTA  
 GCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGT  
 GGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCCTAGGGTTG  
 TAAAGTACTTTTTCAGTCGGGAGGAAGGCGTTGATGCTAATATCATCAACGATTGACGTTACCGACAGAAGA  
 AGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGG  
 CGTAAAGCGCACGCAGGCGGTTGATTAAGTTAGATGTGAAATCCCCGGGCTTAACTGGGAATGGCATCT  
 AAGACTGGTCAGCTAGAGTCTTGTAGAGGGGGGTAGAAATCCATGTGTAGCGGTGAAATGCGTAGAGATG  
 TGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAG  
 CAAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGTCGATTTGAAGGTTGTTCCCTAGAGGA  
 GTGGCTTTCGGAGCTAACGCGTTAAATCGACCGCTGGGGAGTACGGCCGCAAGGTTAAAACCTCAAATGA  
 ATTGACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTAATTCGATGCAACGCGAAGAACCTTACCTAC  
 TCTTGACATCCAGAGAATTTAGCAGAGATGCTTTAGTGCTTCGGGAACTCTGAGACAGGTGCTGCATGG  
 CTGTCGTACGTCGTTGTGTGAAATGTTGGGTTAAGTCCCAGCAGCGCAACCCCTTATCCTTTGTTGC  
 CAGCATTCAGGTCGGGAACTCAAAGGAGACTGCCGTTGATAAACCGGAGGAAGGTGGGGATGACGTCGTAAG  
 TCATCATGGCCCTTACGAGTAGGGCTACACACGTGCTACAATGGCGTATACAAAGAGAAGCGACCTCGCG  
 AGAGCAAGCGGAACCTATAAAGTACGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCCG  
 AATCGCTAGTAATCGTAGATCAGAATGCTACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCA  
 CACCATGGGAGTGGGTTGCAAAGAAGTAGGTAGCTTAACTTCGGGAGGGCGCTTACCACCTTGTGATT  
 CATGACTGGGGTGAAGTCGTAACAAGGTA

>Pseudomonas\_aeruginosa\_FN645737



CACGGATCCAGGACTTTGATTCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTTCGAGC  
GGATGAAGGGAGCTTGTCTCTGGATTTCAGCGCGGACGGGTGAGTAATGCCTAGGAATCTGCCTGGTAGT  
GGGGGATAACGTCCGGAAACGGGCGCTAATACCGCATACTGCTGAGGGGAGAAAGTGGGGGATCTTCGGA  
CCTCACGCTATCAGATGAGCCTAGGTTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGGCGACGAT  
CCGTAACGGTCTGAGAGGATGATCAGTCACACTGGAAGTTCGAGACACGGTCCAGACTCCTACGGGAGGCA  
GCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGTCTTCG  
GATTGTAAAGCACTTAAAGTTGGGAGGAAGGGCAGTAAGTTAATACCTTGCTGTTTTGACGTTACCAACA  
GAATAAGCACCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGTGAAGCGTTAATCGGAATTA  
CTGGGCGTAAAGCGCGCTAGGTGGTTCAGCAAGTTGGATGTGAAATCCCCGGGCTCAACCTGGGAACTG  
CATCCAAAACCTACTGAGCTAGAGTACGGTAGAGGGTGGTGAATTTCTGTGTAGCGGTGAAATGCGTAG  
ATATAGGAAGGAACACCAGTGGCGAAGGCGACCACCTGGACTGATACTGACACTGAGGTGCGAAAGCGTG  
GGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCGTAAACGATGTGCGACTAGCCGTTGGGATCCTT  
GAGATCTTAGTGGCGCAGCTAACGCGATAAGTTCGACCCCTGGGGAGTACGGCCGCAAGGTTAAAACCTCA  
AATGAATTGACGGGGGCCCCGACAAGCGGTGGAGCATGTGGTTTAAATTCGAAGCAACGCGAAGAACCTTA  
CCTGGCCTTGACATGCTGAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTCAGACACAGGTGCTG  
CATGGCTGTGCTCAGCTCGTGTGAGATGTTGGTTAAGTCCCGTAAACGAGCGCAACCCTGTCTCTTA  
GTTACCAGCACCTCGGGTGGGCACTCTAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACG  
TCAAGTCATCATGGCCCTTACGGCCAGGGCTACACACGTGCTACAATGGTTCGGTACAAAGGGTTGCCAAG  
CCGCGAGGTGGAGCTAATCCATAAAAACCGATCGTAGTCCGGATCGCAGTCTGCGACTCGACTGCGTGAA  
GTCGGAATCGCTAGTAATCGTGAATCAGAATGTCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCC  
CGTCACACCATGGGAGTGGGTTGCTCCAGAAGTAGCTAGTCTAACCGCAAGGGGGACGGTTACCACGGAG  
TGATTCATGACTGGGGTGAAGTCGTAACAAGCTAGACCGTAAAGCTTAC

>Rhizobium\_leguminosarum\_HQ218437

TGCAGTCGAGCGCCCCGCAAGGGGAGCGGCAGACGGGTGAGTAACGCGTGGGAATCTACCCTTGACTACG  
GAATAACGCAGGGAACTTGTGCTAATACCGTATGTGTCTTCGGGAGAAAAGATTTATCGGTCAAGGATG  
AGCCCGCTTGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGGCGACGATCCATAGCTGGTCTGAGA  
GGATGATCAGCCACATTTGGGACTGAGACACGGCCCAAACCTCCTACGGGAGGCAGCAGTGGGGAATATTGG  
ACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTGATGAAGGCCCTAGGGTTGTAAAGCTCTTTC  
ACCGGAGAAGATAATGACGGTATCCGGAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGTAATA  
CGAAGGGGGCTAGCCTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGATCAGTCAGGGGT  
GAAATCCCAGGGCTCAACCCTGGAAGTGCCTTTGATACTGTGATCTGGAGTATGGAAGAGGTGAGTGGAA  
ATTCCGAGTGTAGAGGTGAAATTCGTAGATATTCGGAGGAACACCAGTGGCGAAGGCGGCTCACTGGTCC  
ATTACTGACCTGAGGTGCGAAAGCGTGGGAGCAACAGGATTAGATAACCCTGGTAGTCCACGCCGTAA  
ACGATGAATGTTAGCCGTCCGGCAGTATACTGTTGGTGGCGCAGCTAACGCATTAACCATTCGCGCTGG  
GGAGTACGGTTCGCAAGATTAAAACCTCAAAGGAATTGACGGGGGCCCCGACAAGCGGTGGAGCATGTGGTT  
TAATTCGAAGCAACGCGCAGAACCTTACCAGCCCTTGACATGCCCGGCTACTTGCAGAGATGCAAGGTTTC  
CCTTCGGGGACCGGGACACAGGTGCTGCATGGCTGTGCTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTC  
CCGCAACGAGCGCAACCCTCGCCTTAGTTGCCAGCATTAAAGTTGGGCACTCTAAGGGGACTGCCGGTGT  
AAGCCGAGAGGAAGTGGGGATGACGTCAAGTCCATGGCCCTTACGGGCTGGGCTACACACGTGCTAC  
AATGGTGGTGACAGTGGGCAGCGAGCACGCGAGTGTGAGCTAATCTCCAAAAGCCATCTCAGTTCGGATT  
GCACTCTGCAACTCGAGTGCATGAAGTTGGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATAC  
GTTCCCGGGCCTTGTACACACCGCCCGTACACCATGGGAGTTGGTTTTACCCGAAGGTAGTGCCTAAC  
CGCAAGGAGGCAGC

>Serpulina\_hydysenteriae\_M57741

GAGTATGGAGGTTTGGATTCTGGCTCAGAGCGAACGTTGGCGATGCGTCTTAAGCATGCAAGTTCGAGCGG  
GCTTATTCGGGCAACTGGATAAGTTAGCGGCGAACTGGTGAGTAACACGTAGGTAATCTGCCGTAGAGTG  
GGGGATAACCCATGGAAACATGGACTAATACCGCATATACTCTTGCTACATAAGTAGAGTAGAGGAAAGG  
AGCAATCCGCTTTACGATGACNTGCGGCCTATTAGCCTGTTGGTGGGGTAAACGGCCTACCAAAGCTACGA  
TAGGTAGCCGACCTGAGAGGGTGACCGGCCACATTTGGGACTGAGATACGGCCAGACTCCTACGGGAGGC  
AGCAGCTGAGAATCTTCCACAATGGACGAAAGTCTGATGGAGCGACATCGCGTGAGGGATGAAGGCCTTC  
GGGTTGTAACCTCGGAAATATCGAAGAATGAGTGACAGTAGATAATGTAAGCCTCGGCTAACTACGCTG  
CCAGCACCCGNGTNACTAGTAGAGGCAACCTGCTCGGATTTACTGGGCGTAAAGGGTGAAGTAGGCGG  
GACTTATAAGTCTAAGGTGAAAGACCGAAGCTCAACTTCGGAAACGCTCGGATACTGTAAGTCTTGGAT  
ATTGTAGGGGATGATGGAATTTCTCGGTGTAGCGGTGGAATGCGCAGATATCGAGAGGAACANCTATAGCG  
AAGGCAGTCATCTGGGCATTTATCGACGCTGAATCACGAAAGCTAGGGGAGCAACAGGCTTAGATAACC  
TGGTAGTCTTAGCCGTAACGTTGTACACTAGGTGCTTCTATTTAAATAGGAGTGCCGTAGCTAACGCTCT  
TAAGTGTACCGCTGAGGAGTATGCCCGCAAGGGTGAACCTCAAAGAAATTGACGGGTNCCCGCANNAGT  
GGTGGAGCATGTGGTTTAAATTCGATNATACGCGAAAACCTTACCTGGGTTGAATTGTAAGATGAATGAT  
TTAGAGATAAGTCAAGCCGCAAGGACGTTTTACATAGGTGCTGCATGCCTGTGCTCAGCTCGTGTGCTGA

GATGTTGGGTTAAGTCCC<sup>u</sup>GCAACGAGCGCAACCCTCACNNTT<sup>u</sup>GTTGCTACCGAGTAATGTCGGGCACTC  
 TTAGGGGACTGCCTACGTTCAAGTAGGAGGAAGGTGGGGATGATGTCAAGTCCTCATGGCCCTTATGTCC  
 AGGGCTACACACGTGCTACAATGGCAAGTACAAAGAGAAGCAAGACCGCGAGGTGGAGCAA<sup>u</sup>AACTCAAAA  
 AAGTTGCCTCAGTTCCGATTGGAGTCTGAAACTCGACTCCATGAAGTTGGAATCACTAGTAATCGTAGAT  
 CAGAACGCTACGGTGAATACGTTCCCGGNATTGTACACACCGCCCGTACGCCATCGGAGTTNGTTTTA  
 CCTGAAGTCGTTAGCCTAACCGCAAGGGGGCGGCGCCGAAGGTGGNCTGATG

>*Serratia marcescens*\_AY566180

AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGTAGCACAG  
 GGGAGCTTGCTCCCTGGGTGACGAGCGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGG  
 GATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCGGGCCCTC  
 TTGCCATCAGATGTGCCAGATGGGATTAGCTAGTAGTGGGGTAATGGCTCACCTAGGCGACGATCCCT  
 AGCTGGTCTGAGAGGATGACCAGCCACACTGGA<sup>u</sup>ACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAG  
 TGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCCTTCGGGTT  
 GTAAAGCACTTTCAGCGAGGAGGAAGGTGGT<sup>u</sup>GAGCTTAATACGCTCATCAATTGACGTTACTCGCAGAAG  
 AAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGG  
 GCGTAAAGCGCACGACGCGGTTTGTAAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATT  
 TGAAACTGGCAAGCTAGAGTCTCGTAGAGGGGGTAGAATTCAGGTGTAGCGGTGAAATGCGTAGAGAT  
 CTGGAGGAATACCGTGGGCGAAGGCGGCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGG  
 AGCAAACAGGATTAGATA<sup>u</sup>CCCTGGGTAGTCCACGCTGTAAACGATGTGATTTGGGAGGTTGTGCCCTTG  
 AGGCGTGGCTTCGGGAGCTAACCGGTTAAATCGACCGCTGGGGAGTACGGCCGAAGGTTAAACTCAA  
 ATGAATTGACGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTAATTTCGATGCAACGCGAAGAACCTTAC  
 CTACTCTTGACATCCAGAGAACTTTCAGAGATGGATTGGTGCCTTCGGGAACTCTGAGACAGGTGCTGC  
 ATGGCTGTCGTACGCTCGTGTGTGAAATGTGGGTTAAGTCCC<sup>u</sup>GCAACGAGCGCAACCCTTATCCTTTG  
 TTGCCAGCGGTTTCGGCCGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGT  
 CAAGTCATCATGGCCCTTACGAGTAGGGCTACACACGTGCTACAATGGCGTATACAAAGAGAAGCGACCT  
 CGCGAGAGCAAGCGGACCTCATAAAGTACGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAG  
 TCGGAATCGCTAGTAATCGTAGATCAGAATGCTACGGTGAATACGTTCCCGGGCCTTGACACACCGCCC  
 GTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCAC<sup>u</sup>TTTGT  
 GATTCATGACTGGGGTGAAGTCGTAACAAGGTAACC

>*Shewanella colwelliana*\_AY653177

ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGAATTAGCTTGCTAATTTGCTG  
 ACGAGCGCGGACGGGTGAGTAATGCCTAGGGAACGCCAGTCGAGGGGGATAACAGTTGGAAACGACT  
 GCTAATACCGCATA<sup>u</sup>CGCCCTACGGGGGAAAAGAGGGGACCTTCGGGCCCTTCTGCATTGGATGTACATAG  
 GTGGGATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGCGACGATCCCTAGCTGTTCTGAGAGGATGAT  
 CAGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGG  
 GCGCAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCCTTCGGGTTGTAAAGCACTTTCAGCGAGG  
 AGGAAAGCTTAAGCGTTAATAGCGTTTAGGTGTGACGTTACTCGCAGAAGAAGGACCGGCTAACTTCGTG  
 CCAGCAGCCGCGTAATACGAGGGGTCCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGAGGCG  
 GTTTTGTTAAGCGAGATGTGAAAGCCCCGGGCTCAACCTGGGAACTGCATTTTCAACTGGCAA<sup>u</sup>ACTAGAGT  
 CTTGTAGAGGGGGTAGAATTTAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGGTGGCG  
 AAGGCGGCCCTTGGACAAAGACTGACGCTCATGTACGAAAGCGTGGGGAGCAAACAGGATTAGATACCC  
 TGGTAGTCCACGGCTAAACGATGTCTACTCGGAATTTGGTGTCTTGAACACTGGGTTCTCAAGCTAACG  
 CATTAAAGTAGACCGCTGGGGAGTACGGCCGAAGGTTAA<sup>u</sup>AACTCAAATGAATTGACGGGGGCCCGCACA  
 AGCGGTGGAGCATGTGGTTAATTTCGATGCAACGCGAAGAACCTTACCTACTCTTGACATCCAGAGAATT  
 CGCTAGAGATAGCTTAGTGCCTTCGGGAACTCTGAGACAGGTGCTGCATGGCTGTCGTGAGCTCGTGTG  
 TGAAATGTTGGGTTAAGTCCC<sup>u</sup>GCAACGAGCGCAACCCTTATCCTTATTTGCCAGCACGTAATGGTGGGAA  
 CTTTAGGGAGACTGCCGGTGATAAACCGGAGGAAGGTGGGGACGACGTCAGTCATCATGGCCCTTACGA  
 GTAGGGCTACACACGTGCTACAATGGCAAGTACAGAGGGTTGCGAAGCCGCGAGGTGGAGCTAATCTCAC  
 AAAGCTTGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATCGCTAGTAATCGTGG  
 ATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTACACCATGGGAGTGGGCTG  
 CACCAGAAGTAGATAGCTTAACCTTCGGGAGGGCGTTTACCACGGTGTGGTTCATGACTGGGGTGAAGTC  
 GTAACAAGGTAGCCCTAGGGGAACCTGG

>*Skermanella paroensis*\_NR044876

TNTGATCCTGNCTCAGANCGAACGCTGGCGGCATGCCTAACACATGCAAGTCGANCGAGGCCTTCGCCCC  
 TAGTGGCGCACGGGTGAGTAACNCGTGGGAACCTCCCTGTGGTACGGAATAACTCCGGGAAACTGGAGC  
 TAATACCGTATGTGTCTGTGGGACAAAGATTTATCNCCATGGGATGGGCCCGCGTAGGATTAGCTTGT  
 GGTGGGGTAAACGGCTACCAAGGCTTCGATCCTTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACT  
 GAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGCCTGATCCA

ACAATNCCGCGTGAGAGATGAAGGCCTTCGGGTGTAAAGCTCTTTCGCACGCGACGATGATGACGGTAG  
 CGTGAGAAGAAGCCCGGCTAACATCGTGTCCANACNCCGGTAATACGAAGGGGGCTATCGTTCTTCGG  
 AATTACTGGGCGTAAAGGGCGGTAGGCGGTACTTCAAGTCAGGCGTGAAAGCCCGGGCTCAACCCTGG  
 AACCGCGCTTGAGACTGGAGAAGTAGAGTTCGGGAGAGGATGGTGGAAATCCCAGTGTAGAGGTGAAATT  
 CGTAGATATTGGGAAGAACACCGNTGGCGAAGGCAGCCATCTGGACCGACACTGACGCTGAGGCGCGATT  
 GCGTGGGGAGCAAACAGGATTAGATACCCCTGGTAGTCCACGCCGTAACGATGAGTGTAGACGTTGGGG  
 GCCTTAGGCTTCGNTGTTCGACGTAACGCATTAAGCACTCCGCCTGGGGAGTACGGGCGCAAGGTTAAAA  
 CTCAAAGGAATTGACGGGGGCCCGCACAAAGCGGTGGAGCATGTGGTTTTAATTCGAAGCAACGCGCAGAAC  
 CTTACCAGCCCTTGACATGGGCGTTCGGGCTCAGAGATGAGCCTTTCGGTTCGGCCGGACGNCGCACAG  
 GTGCTGCATGGCTGTCTGCTCAGCTCGTGTCTGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCA  
 TCTTCAGTTGGCAGCATGTNATGGTGGGCACTCTGGGAGAACCGCCGGTGACAAGCCGGAGGAAGCGGG  
 GATGACGTCAGTCCCTCATGGCCCTTATGGGCTGGGCTACACACGTGCTACAATGGTGGTGCAGTGGGC  
 AGCGAGATCGCGAGATCGAGCCAATCTCCNAAAGCCATCTCAGTTCGGATCGCACTCTNCAACTCGGGTG  
 CGTGAAGTTGGAATCGCTAGTAATCGCGGATCAGCACNCCCGGTGAATACGTTCCCGGGCCTTGACAC  
 NCCGCCCCGTACACACCATGGAGTTGGTTTTACCCGAAACCGGTGGGCTAACCTCAAGGAGNCAACCGACC  
 ACGGTCAGNTCAACGACTGGGTCC

>Sphingomonas haloaromaticamans\_NR044902

GCTCAGAACGAACGCTGGCGGCATGCCTAACACATGCAAGTCGAACGAAGGCTTCGGCCTTAGTGNCGA  
 CGGGTGCGTAACGCTGGGAATCTNCCCTTGGGTTCGGAATAACAGTGAGAAATTACTGCTAATACCGTA  
 TGATGTCGCGAGACCAAAGATTTATCTCAAAGGATGAGCCCGCTAGGATTAGCTAGTTGGTGGGGTAA  
 AGGCCACCAAGGCGNCGATCCTTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGC  
 CCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGC  
 GTGAGTGAAGGCCCTAGGGTTGTAAAGCTCTTTTACCCGGAAGATAATGACTGTACCGGGAGAATA  
 AGCCCCGCTAACCCGTGCCAGCAGCCGCGTAATACGGAGGGGGCTAGCCTTGTTCGGAATTACTGGG  
 CGTAAAGCGCACGTAGCGGCTTGGTAAAGTTAGAGGTGAAAGCCCGGAGCTCAACTCCGGAATAGCCTTT  
 AAGACTGTCTCGCTTGAACGTCGGAGAGGTGAGTGGAAATCCGAGTGTAGAGGTGAAATTCGTAGATATT  
 CGGAAGAACACCAGTGGCGAAGGCGGCTCACTGGACGACTGTTGACGCTGAGGTGCGAAAGCGTGGGGAG  
 CAAACAGGATTAGATAACCTGGTAGTCCACGCCGTAACGATGATAACTAGCTTGTCCGGGCACTTGTGC  
 TTGGGTGGCGCAGCTAACGCATTAAGTTATCCGCCTGGGAGTACGGTTCGCAAGGTTAAAACCAAAGAA  
 ATTGACGGGGCCTGCACAAGCGGTGGAGCATGTGGTTAATTGCAAGCAACGCGCAGAACCCTTACCAAC  
 GTTTGACATCCCTATCGCGGTAGTGGAGACACTTTCCTTCAGTTCGGCTGGATAGGTGACAGGTGCTGC  
 ATGGCTGTCCGACGCTCGTGTCTGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTCGCCTTTAG  
 TTGCCATCATTAAGTTGGGCACTCTAAAGGAACCCCGGTGATAAGCCGGAGGAAGGTGGGGATGACGTC  
 AAGTCTCATGGCCCTTACGCGTTGGGCTACACAGTGTACAATGGCAACTACAGCAGCGCAACTC  
 GCGAGGGTGAGCTAATCTCCAAAAGTTGTCTCAGTTCGGATTGTTCTCTGCAACTCGAGAGCATGAAGGC  
 GGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCAGGCCTTGTACACACCGCCCGT  
 CACACCATGGGAGTTGGATTACCCGAAGGCGCTGCGCTAACCGCAAGGGAGGCAGGCGACCACGGTGGG  
 TTTAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGG

>Starkeya koreensis\_AB166877

TAAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAACGCACCGCAAGGTGAGTGGCAGACGGGTG  
 AGTAACACGTTGGGATCTGCCAATGGTACGGAATAGCTCCGGGAAACTGGAATTAATACCGTATGTGCC  
 CTTCCGGGGAAAGATTTATCGCCATTGGATGAACCCCGCTCGGATTAGCTAGTTGGTGTGGTAAAGGCGC  
 ACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACCGCCAGAC  
 TCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGT  
 GATGAAGGCCCTTAGGGTTGTAAAGCTCTTTCGCCGACGAAGATAATGACGGTAGTCCGAGAAGAAGCCCC  
 GGCTAACTTCGTGCCAGCAGCCGCGTAATACGAAGGGGGCTAGCGTTGTTCCGGAATCACTGGGCGTAAA  
 GCGCACGTAGGCGGATGTTAAGTCAGGGGTGAAAGCCTGGAGCTCAACTCCAGAACTGCCCTTGATACT  
 GGCAATCTCGAGTCCGGAAGAGGTAAGTGGAACTGCGAGTGTAGAGGTGAAATTCGTAGATATTCGCAAG  
 AACACCAGTGGCGAAGGCGGCTTACTGGTCCGGTACTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAACA  
 GGATTAGATAACCTGGTAGTCCACGCCGTAACGATGGAGGCTAGCCGTTGGTGGAGCATGCTCATCAGTG  
 GCGCAGCTAACGCATTAAGCCTCCCGCTGGGAGTACGGTTCGCAAGATTAACCACTCAAAGGAATTGACG  
 GGGCCCGCACAAGCGGTGGAGCATGTGGTTAATTCGAAGCAACGCGCAGAACCCTTACCAGCCTTTGAC  
 ATGTCCCGAATTGGATCAGAGATGGACCAAGCTCTTCGGAGCCGGGAACACAGGTGCTGCATGGCTGTC  
 GTCAGCTCGTGTCTGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTCGCCCTTAGTTGCCATCA  
 TTCAGTTGGGCACTTAGGGGACTGCCGGTGATAAGCCGAGAGGAAGGTGGGGATGACGTCAGTCAAGTCTC  
 ATGGCCCTTACGGGTGGGCTACACACGTGCTACAATGGCGGTGACAGTGGGAAGCGCAACCCCGAGGGT  
 GAGCAAATCTCCAAAAGCCGTCTCAGTTCGGATTGCATCTGCAACTCGAGTGCATGAAGTTGGAATCGC  
 TAGTAATCGTGGATCAGCACGCCACGGTGAATACGTTCCCGGGCCTTGACACACCGCCCTCACACCAT  
 GGGAGTTGGTTTTACCCGAAAGGCGCTGCGCTAACCCGCAAGGGAGGCAGGCGACCACGGTAGGGTACGCG

ACTGGG

>*Stenotrophomonas maltophilia*\_JN644502

GAACGCTGGCGGTAGGCCTAACACATGCAAGTCAACGGCAGCACAGGAGAGCTTGCTCTCTGGGTGGCG  
 AGTGGCGGACGGGTGAGGAATACATCGGAATCTACTTTTCGTGGGGGATAACGTAGGGAACTTACGCT  
 AATACCGCATACGACCTACGGGTGAAAGCAGGGGACCTTCGGGCCTTGCPCGATTGAATGAGCCGATGTC  
 GGATTAGCTAGTTGGCGGGGTAAAGGCCACCAAGGCAGCATCCGTAGCTGGTCTGAGAGGATGATCAG  
 CCACACTGGAAGTGAACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCG  
 CAAGCCTGATCCAGCCATACCGCGTGGGTGAAGAAGGCCCTTCGGGTTGTAAAGCCCTTTTGTGGGAAAG  
 AAATCCAGCTGGCTAATACCCGGTTGGGATGACGGTACCCAAAGAATAAGCACCGGCTAACCTCGTGCCA  
 GCAGCCGCGGTAATACGAAGGGTGAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCCTAGTGGTCTC  
 GTTTAAGTCCGTTGTAAAGCCCTGGGCTCAACCTGGGAAGTGCAGTGGATACTGGGCGACTAGAGTGTG  
 GTAGAGGGTAGCGGAATTCCTGGTGTAGCAGTGAATGCGTAGAGATCAGGAGGAACATCCATGGCGAAG  
 GCAGCTACCTGGACCAACACTGACACTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGG  
 TAGTCCACGCCCTAAACGATGCGAAGTGGATGTTGGGTGCAATTTGGCACGCAGTATCGAAGCTAACCGG  
 TTAAGTTCCGCCCTGGGAGTACGGTGCAGACTGAACTCAAAGGAATTGACGGGGGCCCGCACAAAG  
 CGGTGGAGTATGTGGTTAATTCGATGCAACGCGAAGAACCTTACCTGGCCTTGACATGTCGAGAATTT  
 CCAGAGATGGATTGGTGCCTTCGGGAAGTCAACACAGGTGCTGCATGGCTGTGCTCAGCTCGTGTGCTG  
 AGATGTTGGGTTAAGTCCCACAACGAGCGCAACCCCTTGTCTTAGTTGCCAGCACGTAATGGTGGGAAGT  
 CTAAGGAGACCGCCGTTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTACGGCC  
 AGGGCTACACACGTAATAAATGGTAGGGACAGAGGGTGAAGCCGGCGACGGTAAGCCATCCAGAA  
 ACCCTATCTCAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATCGCTAGTAATCGCAGAT  
 CAGCATTGCTGCGGTGAATACGTTCCCGGGCCTTGACACACCGCCCGTACACCCATGGGAGTTTGTGTC  
 ACCAGAAGCAGGTAGCTAACCTTCGGGAGGGCGCTTGCCACGGTGTGGCCGATGACTGGGGTGAAGTCG  
 TACAAG

>*Thioreductor micantisoli*\_NR041022

AGTGAACGCTGGCGCATGCCTAACACATGCAAGTCAACGAGAACGGATCTAACTTCGGTTAGATTGTC  
 AGCTAAGTGGCGCACGGGTGAGTAACACGTAGTTATCTGCCTCACAGTCTGGGATAACAATTGGAAACGA  
 TTGCTAATACCGGATATACCCACGGGGGAAAGCTTTAGTGTGTGAGATGAGACTGCGACGTATCAGCT  
 AGTTGGTAGGGTAAGAGCCTACCAAGGCTAAGACGCGTAACTGGTCTGAGAGGATGATCAGTCACACTGG  
 AACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGAGGAACTCTGA  
 TGCAGCAATGCCGCGTGGAGGATGACGGATTTCCGGTCTGTAAACTCCTTTTATATAGGAAGATAATGACG  
 GTACTATATGAATAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGAACGCGTTAC  
 TCGGAATCAGTGGGCGTAAAGAGCGCGTAGGCTGGTTTGTAAAGTTAGAAGTGAATCCCCAGCTCAACT  
 GTGGAAGTGCCTTCTAAACTGCAGACCTAGAATTTGGGAGAGGTAAGTGAATTCCTGGTGTAGGGGTGA  
 AATCCGTAGAGATCAGGAGGAATACCGAAAGCGAAGGCGACTTACTGGAACAATATTGACGCTGATGCGC  
 GAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGGACACTAGTGCCT  
 GCTATAAAGCAGTACGACGCTAACGCGATAAGTGTCCCGCTGGGGAGTACGGTGCAGGATTTAAACT  
 CAAAGGAATAGACGGGACCCGAACAAGCGGTGGAGCATGTGGTTAATTCGAAGATACGCGAAGAACCT  
 TACCTAGGCTTGACATCCCAAGAATCTTTAGAGATGAGAGAGTGCCTGCTTGCAGGAACTTGGTGACAG  
 GTGCTGCACGGCTGTGCTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTG  
 TTGTTAGTTGCTAACAGTTCGGCTGAGCACTCTAACAGACTGCCTGGGTAACCAGGAGGAAGGTGGGGA  
 CGACGTCAGTATCATGAGCCCTTATGCCTAGGGCGACACACGTCATCAATGGTTAGGATAAAGAGACG  
 CAATACCGCGAGGTGAGCAAATCTCTAAACCTAATCCAGTTCGGATTGTAGTCTGCAACTCGACTACA  
 TGAAGTTGGAATCGCTAGTAATCGTGGATCAGCCATGCCACGGTGAATACGTTCCCGGGTCTTGTACTCA  
 CCGCCCGTACACCATGGGAGTTGAGTTCACCCAAAGCGGGGATGCCAAATTGGCTACCCTCTACGGTGG  
 GCTCAGCAACTGGGGTG

>*Treponema denticola*\_AR621358

AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCGCGTCTTAAGCATGCAAGTCAACGGTAAGGGAG  
 AGCTTGCTCTCCCTAGAGTGGCGGACTGGTGAAGTACCGGTGGGTGACCTGCCCTGAAGATGGGGATAG  
 CTAGTAGAATAATTAGATAAATACCGAATGTGCTCAATTTACATAAAGGTAATGAGGAAAGGAGCTACGGC  
 TCCGCTTCAAGGATGGGCCCGCTCCCATAGCTGTTGGTGAAGTAAAGGCCACCAAGGCAACGATGGG  
 TATCCGGCCTGAGAGGGTGAACGGACACATTTGGGACTGAGATACGGCCAAACTCCTACGGGAGCGACGA  
 GCTAAGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCGACGCCGTGTAATGAAGAAGGCCGAAAGGT  
 TGTAATAATCTTTTGCAGATGAAGAATAAGAAGAAGAGGGAATGCTTCTTTGATGACGGTAGTCATGCGA  
 ATAAGCCCCGGCTAATTACGTGCCAGCAGCCCGGTAACACGTAAGGGGCGAGCGTTGTTCCGAATTTAT  
 GGGCGTAAAGGGTATGTAGGCGGTTAGGTAAGCCTGGTGTGAAATCTACGAGCTCAACTCGTAAACTGCA  
 TTGGGTACTGCTTGAATCACGGAGGGGAAACCGGAATCCAAGTGTAGGGGTGGAATCTGTAGAT  
 ATTTGGAAGAACCAGGTGGCGAAGGCGGGTTTCTGGCCGATGATTGACGCTGATATACGAAGGTGCGGG

GAGCAAACAGGATTAGATACCCTGGTAGTCCGCACAGTAAACGATGTACACTAGGTGTCGGGGCAAGAGC  
TTCGGTGCCGACGCAAACGCATTAAGTGTACCGCCTGGGAAGTATGCCCGCAAGGGTGAAACTCAAAGGA  
ATTGACGGGGGCCACACAAGCGGTGGAGCATGTGGTTAATTCGATGATACGCGAGAAACCTTACCTGG  
GTTTGACATCAAGAGCAATGACATAGAGATATGGCAGCGTAGCAATACGGCTCTTGACAGGTGCTGCATG  
GCTGTCGTCAGCTCGTGCCGTGAGGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTACTGCCAGTTA  
CTAACAGGTAAAGCTGAGGACTCTGGCGGAACTGCCGATGACAAATCGGAGGAAGGTGGGGATGACGTCA  
AGTCATCATGGCCCTTACGTCCAGGGCTACACACGTGCTACAATGGTTGCTACAAATCGAAGCGACACCG  
CGAGGTCAAGCAAACGCAAAAAAGCAATCGTAGTCCGGATTGAAGTCTGAAACTCGACTTCATGAAGTT  
GGAATCGCTAGTAATCGCACATCAGCACGGTGCGGTGAATACGTTCCCTGGGCCTTGACACACCGCCCGT  
CACACCATCCGAGTCGAGGGTACCCGAAGTCGCTAGTCTAACCCGTAAGGGAGGACGGTGCCGAAGGTAC  
GTTTGGTAAGGAGGGTGAAGTCGTAACAAGGTAACC

## Appendix B

### Amino acid sequences of glycoside hydrolase family 45

These are 5 amino acid sequences of species in Glycoside Hydrolase Family 45. All are listed in FASTA format. Accession number of each species id located at the end of each species name.

```
>Cellvibrio_japonicus_ACE82688
MNLISGWRPLMLGCGLLGAALSAGSIQAAVCEYRVTNEWGSGFTASIRITNNGSSTINGWSVSWNYTDG
SRVTSSWNAGLSGANPYSATPVGWNTSIPIGSSVEFGVQGNNGSSRAQVPAVTGAICGGQGSAPSSVAS
SSSSSSVSSSTPRSSSSSVSSSVPGTSSSSSSSVLTGAQACNWTGTLTPLCNNTSNGWGYEDGRSCVART
TCSAQAPYGIIVSTSSSTPLSSSSSSRSSVASSSSLSATSSSASSVSSVPPIDGGCNGYATRYWDCKP
HCGWSANVPSLVSPQLQSCSANTRLSDVSVGSSCDGGGGYMCWDKIPFAVSPTLAYGYAATSSGDVCGRC
YQLQFTGSSYNAPGDPGSAALAGKTMIVQATNIGYDVSGGQFDILVPGGGVGAFNACSAQWGVSNALGA
QYGGFLAACKQQLGYNASLSQYKSCVLNRCDSVFGSRGLTQLQQGCTWFAEFEAADNPSLKYKEVPCPA
ELTTRSGMNRSLNDIRNTCP
```

```
>Fusarium_oxysporum_AAA65589
MRSYTLALAGPLAVSAASGSGHSTRYWDCKPSCSWGKAAVNAPALTCDKNDNPISTNAVNGCEGGG
SAYACTNYSWAVNDELAYGFAATKISGGSEASWCCACYALTFTTGPVKGKKMIVQSTNTGGDLGDNHFD
LMMPGGGVGIFDGTSEFGKALGGAQYGGISSRSECDYPELLKDGCHWRFDWFENADNPDFTFEQVQCP
KALLDISGCKRDDSSFFPAFKGDTASAKPQPSSAKKTTSAAAAQPKTKDSAPVVQKSSTKPAAQPEP
TKPADKPKQTDKPVATKPAATKPAQPVNKPKTTQKVRGKTRGSCPAKTDATAKASVVPAYYQCGGSKSAY
PNGNLACATGSKCVKQNEYYSQCVPN
```

```
>Humicola_grisea_AAE55435
MRSSPLLRSAVVAALPVLALAADGKSTRYWDCKPSCGWAKKAPVNQPVFSCNANFQRLTDFDAKSGCEP
GGVAYSCADQTPWAVNDDFAFGFAATSIAGSNEAGWCCACYELTFTSGPVAGKKMVVQSTSTGGDLGNSH
FDLNIPIGGGVGIFDGTQFQGLPGQRYGGISSRNECDRFPDALKPGCYWRFDFWKNADNPSFSFRQVQC
PAELVARTGCRNDDGNFPAVQIPSSSTSSPVGQPTSTSTSTSTSTSSPPVQPTTPSGCTAERWAQCGGN
GWSGCTTCVAGSTCTKINDWYHQCL
```

```
>Humicola_grisea_BAA74957
MQLPLTLLTLLPALAAAQSGSRTTRYWDCKPSCAWPGKGPAPVVRTCDRWDNPLFDGGNTRSGCDAGG
GAYMCSQSPWAVSDDLAYGWAAVNIAGSNERQWCCACYELTFTSGPVAGKRMIVQASNTGGDLGNNHFD
IAMPGGGVGIFNACTDQYGAPPNGWQRYGGISQRHECDAFPEKLLKPGCYWRFDFWCVSLFPPLSLSLPPG
TGQTMGRSCVFFPLSAN
```

```
>Humicola_insolens_AAE16508
MRSSPLLPASVVAALPVLALAADGRSTRYWDCKPSCGWAKKAPVNQPVFSCNANFQRLTDFDAKSGCEP
GGVAYSCADQTPWAVNDDFALGFAATSIAGSNEAGWCCACYELTFTSGPVAGKKMVVQSTSTGGDLGNSH
FDLNIPIGGGVGIFDGTQFQGLPGQRYGGISSRNECDRFPDALKPGCYWRFDFWKNADNPSFSFRQVQC
PAELVARTGCRNDDGNFPAVQIPSSSTSSPVNQPTSTSTSTSTSTSSPPVQPTTPSGCTAERWAQCGGN
GWSGCTTCVAGSTCTKINDWYHQCL
```

## Appendix C

### Installation of CUPrimer program.

In fact CUPrimer program is .jar program as the portable version that no need to install on your desktop computer. But before use the CUPrimer program we need to prepare the system which is Java Runtime Environment, is also referred to as the Java Runtime, Runtime Environment, Runtime, JRE, Java Virtual Machine, Virtual Machine, Java VM, JVM, VM, or Java download.

1. Download the file by go to URL: <http://java.com/en/download/index.jsp> and click **Free Java Download**.



2. System will automatically detect installed Java version on your computer.



3. System will select the suitable version for your desktop computer and click **Agree and Start Free Download.**





4. Click on **Run** button to start the installation process.



5. Click **Install** and wait for 2-3 minutes for download and installation process, after finished installation, click close button to exit the program



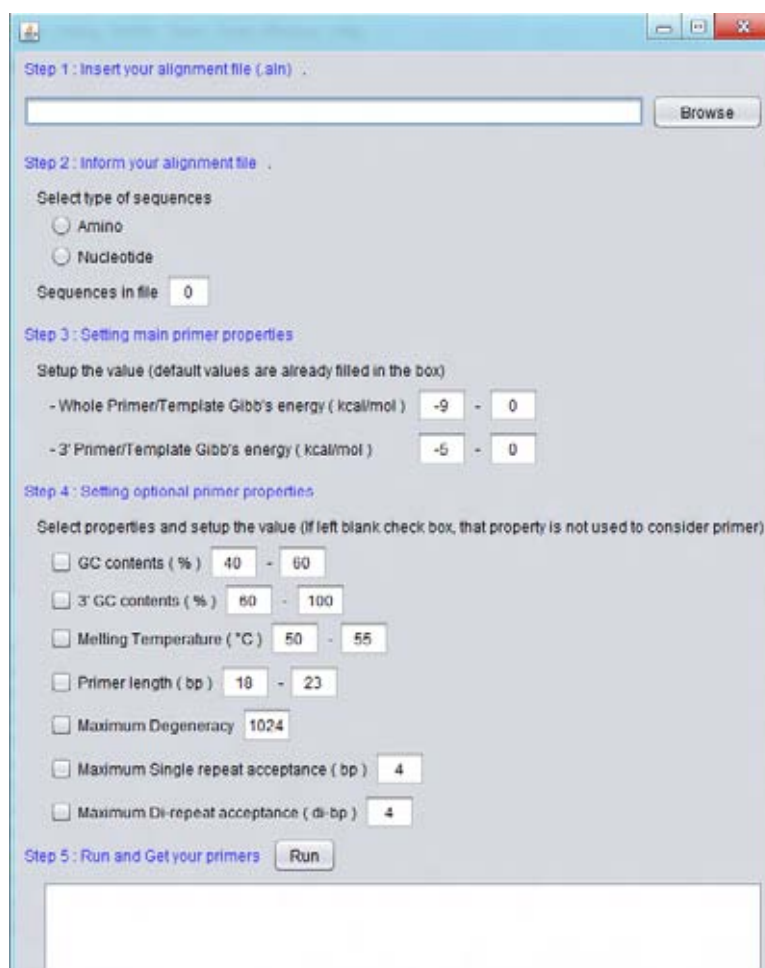


## Appendix D

### CUPrimer program manual guide.

CUPrimer program manual is separate into 5 main steps; input alignment file, inform alignment file, set main primer properties, set optional primer properties and program running step.

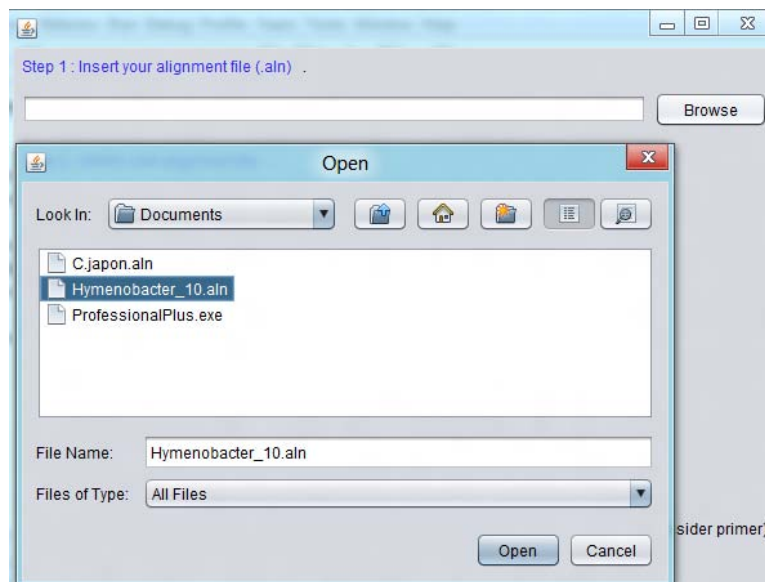
1. Double click CUPrimer.jar to open the program.



The screenshot displays the CUPrimer program interface, which is organized into five sequential steps:

- Step 1: Insert your alignment file (.aln)**: A text input field is provided for the alignment file, accompanied by a "Browse" button.
- Step 2: Inform your alignment file**: This step includes a "Select type of sequences" section with radio buttons for "Amino" and "Nucleotide". Below this, there is a "Sequences in file" input field with the value "0".
- Step 3: Setting main primer properties**: This step instructs the user to "Setup the value (default values are already filled in the box)". It features two rows of energy settings: "- Whole Primer/Template Gibb's energy ( kcal/mol )" with values "-9" and "0", and "- 3' Primer/Template Gibb's energy ( kcal/mol )" with values "-5" and "0".
- Step 4: Setting optional primer properties**: This step instructs the user to "Select properties and setup the value (if left blank check box, that property is not used to consider primer)". It lists several optional properties, each with a checkbox and a value range:
  - GC contents ( % ): 40 - 60
  - 3' GC contents ( % ): 60 - 100
  - Melting Temperature ( °C ): 50 - 55
  - Primer length ( bp ): 10 - 23
  - Maximum Degeneracy: 1024
  - Maximum Single repeat acceptance ( bp ): 4
  - Maximum Di-repeat acceptance ( di-bp ): 4
- Step 5: Run and Get your primers**: A "Run" button is located at the bottom of this step, above a large empty text area for the output.

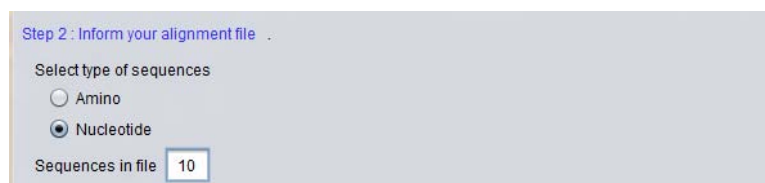
2. Click **Browse** button and choose the alignment file that you want to design primer.



3. After choosing the alignment file it will show the file name on the text box.



4. Select sequence type and enter amount of sequence in alignment file.





## BIOGRAPHY

Weeris Treeratanajaru was born at Bangkok, Thailand. He received a bachelor degree from Environmental Science Program, Faculty of Science, Chulalongkorn University in 2010. Now he is studying a Master Degree in Computer Science and Information from Chulalongkorn University, and planning to qualify a doctorate degree in Computer Science too.

Research paper

1. "Degenerate Primer Designing System for Gene Biodiversity Study Using Dynamic Pattern Matching" presented at Nevsehir, Turkey and published in IEEE-Computer Society.