COMPRESSION OF DNA SEQUENCES

THEERACHAI LAOKULSANT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2002

ISBN 974-04-2320-5

4037517 SCCS/M : MAJOR: COMPUTER SCIENCE; M.Sc. (COMPUTER SCIENCE)
KEY WORDS : COMPRESSION/ DNA COMPRESSION
THEERACHAI LAOKULSANT : COMPRESSION OF DNA SEQUENCES. THESIS ADVISORS: DAMRAS WONGSAWANG Ph.D., SUKANYA PHONGSUPHAP Ph.D., 215 P. ISBN 974-04-2320-5.

Nowadays, there is an evolution of decoding a genetic code of the deoxyribonucleic acid (DNA) and there is a prediction that the 21st century will be a century of biotechnology. DNA sequences can be considered as texts over a four-letter alphabet {(A) adenine, (C) cytosine, (G) guanine, (T) thymine}. For complete genomes, these texts can be very long. The human genome, for instance, contains three billion characters over twenty-three pairs of chromosomes. The fact described above illustrates that decoding a genetic code of DNA needs a huge storage space to store the DNA data. One approach that helps to decrease this needed storage space is data compression. Unfortunately, the compression of DNA sequences appears to be a difficult task. They are, at first glance, very similar to random string, and have only very hidden regularities. The classical algorithms for a text compression do not work on DNA sequences.

This research aimed to propose three approaches of the DNA sequence compression that consume a storage space of fewer than two bits for each character. IbioCompress algorithm is improved from Biocompress-2 with a combination of the techniques of LZ77 and LZ78 algorithms. It has three search matching methods; namely, forward, reverse and inverse search matching methods, which have no boundary of search matching (sliding windows of no limited size). These methods use two bits for encoding literal words. IbioCompress also uses other methods for encoding; namely, adaptive arithmetic order-1, order-2 or PPMZ-M. TCompress algorithm is the approach for transforming the DNA sequence data in an ACGT format and then into many other formats; namely, C4C, C3C, C2C and C01, before compressing it (in ACGT, C4C, C3C, C2C, and C01 format) with good and popular compression programs, like Winzip, Pkzip, Gzip, PPMZ, BOA, Rk, etc. Finally, IgenCompress algorithm is improved from GenCompress by modifying the encoding edit operation and encoding literal words using adaptive arithmetic order-1 or PPMZ-M by comparison and encoding with the best.

As a result, the algorithms of TCompress and IgenCompress consume a storage space of fewer than 2 bits for each character. With the TCompress algorithm, the DNA sequences in C4C format can also be compressed with other classical algorithms for a text compression, like Winzip, Pkzip and Gzip. The DNA sequences in an ACGT format can be compressed with PPMZ algorithm with the best compression ratio. Finally, it can be concluded that in general, IgenCompress algorithm has a better compression ratio than the algorithms of TCompress, Biocompress-2 and GenCompress.

4037517 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์; วท.ม. (วิทยาการคอมพิวเตอร์)

ธีรชัย เลากุลศานต์ : การบีบอัดลำดับข้อมูลของพันธุกรรม (COMPRESSION OF DNA SEQUENCES). คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D., สุกัญญา พงษ์สุภาพ, Ph.D., 215 หน้า. ISBN 974-04-2320-5.

ปัจจุบัน การศึกษาเรื่อง รหัสพันธุกรรม (DNA : Deoxyribo Nucleic Acid) ได้รับความสนใจและเจริญกว้าหน้ามากขึ้นทุกวัน จนสามารถทำนายได้ว่า ศตวรรษที่ 21 นี้จะเป็น ศตวรรษแห่ง เทคโนชีวภาพ รหัสพันธุกรรมนี้ จะมีส่วนประกอบพื้นฐานจำนวน 4 ตัวซึ่งสามารถแทนได้ด้วยตัวอักษร {(A) adenine, (C) cytosine, (G) guanine, (T) thymine}. คาดการว่า การถอดรหัสพันธุกรรมของมนุษย์ จะได้ตัวอักษรดังกล่าวเรียงตัวกันยาวประมาณ 3 พันล้านตัวอักษร จาก 23 คู่ของโครโมโซม ซึ่งจะเสียเนื้อที่ในการจัดเก็บเป็นจำนวนมาก วิธีในการลดการใช้พื้นในการจัดเก็บก็คือ การบีบอัดข้อมูลนั้นเอง แต่ก็มีปัญหา เพราะไม่สามารถทำการบีบอัดด้วยโปรแกรม ที่นิยมใช้ในการบีบอัดทั่ว ๆ ไปได้ เช่น Winzip, Pkzip, Gzip เป็นต้น เนื่องจากอัดลำดับข้อมูลของรหัสพันธุกรรมจะค่อนข้างเรียงตัวกันแบบสุ่ม ไร้กฎเกณฑ์ที่แน่นอน ไม่สามารถคาดการณ์ได้

งานวิจัยฉบับนี้ได้ทำการศึกษาและเสนอแนวทางในการบีบอัดอัดลำดับข้อมูลรหัสพันธุกรรม ให้ใช้เนื้อที่ในการจัดเก็บที่น้อยกว่า 2 บิตต่อตัวอักษร ซึ่งได้ทำการเสนอแนวทางไว้ 3 แนวทางด้วยกันคือ IbioCompress, TCompress, และ IgenCompress

IbioCompress เป็นการปรับปรุงจาก Biocompress-2 โดยใช้เทคนิคของ LZ77 และ LZ78 ร่วมกัน โดยมีรูปแบบการค้นหาถึง 3 รูปแบบด้วยกัน (forward, reverse, inverse) และขอบเขตในการค้นหาจะเป็นแบบเปิด (sliding windows : no limited). และทำการ encode literal word ด้วย 2 บิต หรือ ด้วยวิธีอื่น (arithmetic order-1, order-2, PPMZ-M)

TCompress เป็นการแปลงข้อมูลของรหัสพันธุกรรมไปเป็นรูปแบบต่าง ๆ (ACGT, C4C, C3C, C2C, C01) แล้วทำการบีบอัดด้วยโปรแกรม ที่นิยมและปัจจุบันใช้กันทั่ว ๆ ไป เช่น Winzip, Pkzip, Gzip, PPMZ, BOA, Rk เป็นต้น

IgenCompress เป็นการปรับปรุงจาก GenCompress โดยทำการปรับปรุงในส่วนของ encoding edit operation และในส่วนของ encoding literal word โดยนำเอา adaptive arithmetic order-1 และ PPMZ-M มาใช้ร่วมกัน โดยทำการเปรียบเทียบแล้วทำการบีบอัดด้วยตัวที่ดีที่สุด

ผลจากการทดลอง TCompress และ IgenCompress ใช้เนื้อที่ในการจัดเก็บน้อยกว่า 2 บิตต่อตัวอักษร โดยใน TCompress รูปแบบ C4C สามารถถูกบีบอัดด้วยโปรแกรม ที่นิยมใช้ในการบีบอัดทั่ว ๆ ไป แต่ในรูปแบบ ACGT ที่ถูกบีบอัดด้วย PPMZ จะใช้เนื้อที่ในการจัดเก็บน้อยที่สุด และ IgenCompress จะใช้เนื้อที่ในการจัดเก็บน้อยที่สุด คือน้อยกว่า TCompress, Biocompress-2 และ GenCompress.