บทคัดย่อภาษาไทย

นับตั้งแต่การประกาศเริ่มต้นโครงการจีโนมมนุษย์ในปี ค.ศ. 1990 จนกระทั่งการถอดรหัสโครโมโซมทั้งหมดเสร็จสมบูรณ์ในปี ค.ศ. 2003 ทำให้เกิดการถอดรหัสเซลล์ของสิ่งมีชีวิตและสร้างข้อมูลจีโนมเพิ่มขึ้นอย่างรวดเร็วมากในแต่ละปี แต่การตีความเพื่อใช้ประโยชน์ข้อมูลจีโนมเหล่านี้ทำได้ช้ามาก ทำให้เกิดงานด้านชีวสารสนเทศศาสตร์ ที่เน้นการนำเทคโนโลยีคอมพิวเตอร์มาช่วยในการรู้จำรหัสพันธุกรรมและการชี้ตำแหน่งโครงสร้างหลักในสายดีเอ็นเอ เช่นการระบุตำแหน่งผู้ให้และผู้รับในสายดีเอ็นเอที่เป็นขั้นตอนเริ่มต้นของการสร้างโปรตีน ขั้นตอนวิธีที่มักจะใช้ในการรู้จำรหัสพันธุกรรมประกอบด้วยเทคนิคฮิดเดนมาร์คอฟโมเดล โครงข่ายแบบเบย์ส์ และกำหนดการพลวัต ในระยะหลังเริ่มมีการใช้เทคนิคอัจฉริยะเช่น โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน และจีเนติกอัลกอริทึมมาช่วยเพิ่มความแม่นยำของการรู้จำรหัสพันธุกรรม

ในโครงการวิจัยนี้ผู้วิจัยกำหนดขอบเขตการศึกษาการรู้จำตำแหน่งพันธุกรรมในสายดีเอ็นเอที่เกี่ยวข้องกับการสังเคราะห์โปรตีนของสิ่งมีชีวิตชั้นสูง การรู้จำในงานวิจัยนี้เน้นการจำแนกตำแหน่งเชื่อมต่อระหว่างส่วนอินทรอนและเอ็กซอนในสายดีเอ็นเอขนาดสั้น เทคนิคการรู้จำเป็นการประยุกต์กระบวนการวิศวกรรมความรู้ร่วมกับขั้นตอนวิธีทางปัญญาประดิษฐ์ โดยในกระบวนการดังกล่าวจะรวมขั้นตอนการคัดเลือกฟีเจอร์ การสร้างโมเดล การประเมินความถูกต้องของโมเดล และการแปลงโมเดลเป็นกฎเพื่อแสดงผลลัพธ์ในลักษณะของฐานความรู้ การนำเสนอโมเดลในรูปแบบของฐานความรู้จะช่วยให้ผู้ใช้งานโปรแกรมทำความเข้าใจได้ง่าย นอกจากนี้ยังช่วยให้ปรับเปลี่ยนโมเดลได้สะดวกเมื่อข้อมูลในอนาคตมีการเปลี่ยนแปลง โปรแกรมที่ออกแบบยังสามารถทำงานในลักษณะยืดหยุ่นในกรณีที่ข้อมูลมีความไม่สมบูรณ์ทำให้โมเดลเป็นการค้นหาความรู้โดยประมาณ การพัฒนาโปรแกรมของงานวิจัยนี้ใช้ภาษาเออแลง และเขียนโปรแกรมให้ทำงานได้ทั้งในแบบการโปรแกรมแบบลำดับและการโปรแกรมแบบขนาน โปรแกรมทั้งสองแบบนี้เปิดเผยซอร์สโค้ดเพื่อให้นักวิจัยที่สนใจสามารถพัฒนาต่อยอดงานวิจัยได้

บทคัดย่อภาษาอังกฤษ

Since the announcement of the human genome project in 1990 up to the successful sequencing of all the human chromosomes in the year 2003, the amount of available genome data has been increasing exponentially each year. Unfortunately, genomic interpretation cannot keep pace with such tremendous raw sequenced data. Computational methods to gene recognition and identification of its structural elements such as donor and acceptor splice sites are thus important to the success of bioinformatics. The widely used methods for gene recognition include hidden Markov model, Bayesian network, and dynamic programming. Recent advances in gene prediction tools apply computational intelligent methods such as artificial neural network, support vector machines, and genetic algorithms to produce a more accurate model.

In this project, we consider the problem of recognizing coding regions for protein biosynthesis in eukaryotes. The recognition task is to separate coding and non-coding regions, and to identify the boundaries of intron and exon parts in the unknown DNA sequences. We tackle the problem with the knowledge engineering approach in which not only the machine learning techniques are employed, but also the whole process of knowledge discovery including feature selection, data modeling, model validation, and rule extraction is to be designed and developed. The advantages of the proposed knowledge engineering approach are the ease of use, the automatic generation of informative and comprehensible model, and the adaptation on new information. The induced prediction model is also expected to work well with approximate, incomplete, and uncertain data due to ambiguity in DNA sequencing. We developed the DNA coding region recognition program with the Erlang programming language in both sequential and parallel modes. The program is open source in such a way that the source code is publicly available for further improvement by interesting researchers.