วัตถุประสงค์ของการวิจัยครั้งนี้คือเพื่อศึกษาการจัดกลุ่มโดยใช้สถิติแบบเบส์ ที่อาศัยการแจกแจงก่อนและฟังก์ชันควรจะเป็นเพื่อหาการแจกแจงภายหลังของพารามิเตอร์ที่สนใจ สำหรับการจัดกลุ่มแบบเบส์จะเพิ่มพารามิเตอร์บ่งชี้ในตัวแบบเพื่อกำหนดกลุ่มให้แก่ข้อมูลแต่ละตัว โดยทำการประยุกต์การจัดกลุ่มแบบเบส์ร่วมกับการสุ่มตัวอย่างแบบกิบส์เพื่อจัดกลุ่ม Transcription Factor Binding Site (TFBS) ที่อยู่บริเวณ Promoter ของยีนที่ถูกควบคุมในชุดข้อมูลจีโนม ซึ่งการรู้ถึงกลุ่มยีนที่ถูกควบคุมและกลุ่ม TFBS เป็นวิธีหนึ่งที่จะช่วยให้เข้าใจการควบคุมการแสดงออกทางพันธุกรรมได้

เพื่อทำการประเมินผลการจัดกลุ่มแบบเบส์โดยเปรียบเทียบกับวิธีการจัดกลุ่มแบบ Hierarchical และวิธีการจัดกลุ่มแบบ K-mean โดยอาศัยชุดข้อมูลที่จำลองจาก TFBS ที่ได้จากจีโนมของการศึกษาวิวัฒนาการของแบคทีเรีย *Escherichia coli* จากกลุ่ม TFBS ที่มีอยู่ทั้งหมด 43 กลุ่ม ทำการจำลองกลุ่ม TFBS ขึ้นมาให้มีจำนวนกลุ่มต่าง ๆ กัน และในแต่ละกลุ่ม TFBS ทำการจำลองให้มีขนาดต่าง ๆ กันออกไป ในการเปรียบเทียบผลการจัดกลุ่มระหว่างวิธีการจัดกลุ่มทั้ง 3 วิธีจะพิจารณาจากค่าเฉลี่ยร้อยละของการจัดกลุ่มผิดพลาด ผลการศึกษาพบว่าการจัดกลุ่มแบบเบส์จะให้ผลการจัดกลุ่มที่ดีที่สุดในกรณีที่กำหนดจำนวนและขนาดของกลุ่มอย่างสุ่มจากช่วงที่แตกต่างกันและในกรณีที่จำนวนกลุ่มคงที่เท่ากับ 5 หรือมากกว่า สำหรับกรณีที่จำนวนกลุ่มคงที่ที่น้อยกว่า 5 การจัดกลุ่มแบบ Hierarchical ให้ความผิดพลาดในการจัดกลุ่มต่ำกว่าการจัดกลุ่มแบบเบส์เพียงเล็กน้อย จึงสรุปได้ว่าการจัดกลุ่มแบบเบส์ร่วมกับการสุ่มตัวอย่างแบบกิบส์มีความพอเพียงและเหมาะสมที่จะนำไปใช้ประโยชน์ในการจัดกลุ่ม TFBS ของสปีชีส์อื่นซึ่งทราบผลการจัดกลุ่มด้วยวิธีการทดลองค่อนข้างน้อย

218213

The objective of this research is to study the clustering using Bayesian approach which uses the prior distribution and likelihood function to find the posterior distribution of the interest parameters. In Bayesian clustering, the indicator parameter is added into the model to assign cluster number to each case. The Bayesian clustering combined with the Gibbs sampling method were applied to cluster the Transcription Factor Binding Site (TFBS) in promoter regions of coexpressed genes based on genome sequence dataset. The identification of co-regulated genes and their TFBS are key steps toward understanding transcription regulation.

In order to evaluate the performance of Bayesian clustering, we compared it with Hierarchical clustering and K-mean clustering methods, on simulated data sets of TFBS or motif derived from a genome-scale phylogenetic footprinting study of *Escherichia coli* TFBS. We generated a varied number of motif from 43 distinct regulons or clusters. Each motif consisted of a varied size. The overall performance of each clustering method was assessed by the percent clustering errors averaged. The results indicated that Bayesian clustering yielded the best clustering results when the number of motif and each motif size were randomized over different range or fixed as equal or more than 5. Hierarchical clustering produced a little fewer errors than Bayesian clustering when the number of motif and each motif size were fixed as less than 5. In conclusion, Bayesian clustering and Gibbs sampling implementation are general enough to be appropriate and useful in clustering motif patterns of other species where few regulons have been experimentally determined.