סיווג קטעי בקרה מהגנום האנושי על ידי מודל מרקוב חבוי מוכלל מסדר גבוה

מגיש:
דוד טאוב

מנחה:
פרופ' תומי קפלן

עבודת גמר לתואר מוסמך מדעי המחשב

מאי 2020

בית הספר להנדסה ולמדעי המחשב על שם רחל וסלים בנין
האוניברסיטה העברית בירושלים, ישראל

תקציר

מעצמים (enhancers) הם קטעי בקרה גנטיים אשר מגבירים את סיכויי השעתוק של גן המטרה שלהם על ידי קשירה לחלבונים גורמי שעתוק (transcription factors) והצמדות לאזור הגן. בקרת השעתוק הינה צורת שליטה חשובה בביטוי הגנים וממלאת תפקיד משמעותי בבקרת הגנים על פי מצב התא והרקמה לה הוא שייך. במשך השנים נצטברו עדויות לכך ששינויים גנטיים ברצפי המעצמים עלולים לגרום לשינויים בהתנהגותם של תאים וכתוצאה מכך, למחלות. הכללים והדקויות של מבנה המעצם טרם הובנו לחלוטין, אף כי ניסויים הראו כי גורמי שעתוק נוטים להצמד לאתרי קשירה של גורמי שעתוק, שהם מוטיבים יחודיים אשר נפוצים יחסית בקטעי המעצמים. ניתן לאתר את פעילות המעצם לפי מידע אפיגנטי בסביבתו, כשהסימנים העיקריים הם השינויי בהיסטונים (histone modifications) אשר סביבם כרוכים אגפי המעצם. המעצם נוטה להיות נגיש מרחבית עבור אינטראקציות ביוכימיות בין ה־DNA לחלבונים שסביבו. על אף תועלתו הרבה, המידע האפיגנטי לעיתים קרובות רועש ומצריך תהליכים יקרים של חילוץ תאים ספציפיים מתוך רקמותיהם, פעולה אשר אינה תמיד ברת ביצוע לכלל סוגי התאים בשלביהם השונים. דרך נוספת לזיהוים של מעצמים היא בחינת תוכנם של ריצפם הגנטי, משום שבהם נמצא כל המידע הדרוש ל־DNA על מנת שיתחיל לפעול כמעצם. ניסויים אשר נערכו בחיות מעבדה הראו כי התא איננו זקוק למנגנון נוסף מעבר לרצף הגנטי על מנת לגרום לבקרת הגנים שבו. בשל תפישה זו, אנו מציעים גישה חישובית לזיהוי של מעצמים על בסיס ריצפם הגנטי בלבד, בדרך למידה לא מונחית. יצרנו מודל מרקוב חבוי מסדר גבוה מבוסס מטריצות משקול מיקום, בעל שני סוגי מצבים: מצב אחד אשר פולט אתרי קשירה של גורמי שעתוק מתוך מטריצות משקול מקום, ואחד אשר פולט נוקלאוטידים בודדים תוך תלות מסדר גבוה באלה שנפלטו לפניהם. בהשוואה למודל מרקוב חבוי רגיל, מודל זה לומד מבנה מורכב יותר של הרצף הגנטי, אשר מכיל מוטיבים של אתרי קשירה והתפלגות מסדר גבוה של נוקלאוטידים המצויים בינהם. אנו נבחן תחילה את הרקע הביולוגי של המעצמים, תוך התרכזות בבני אדם. לאחר מכן נסקור לעומק את הרקע של מודלי מרקוב חבויים, ונדון בדרך לחישוב הנראות (likelihood) של רצפים בהינתן המודל. נתאר לפרטים את המודל המוכלל שלנו ונפתח את אלגוריתם מיקסום התוחלת (expectation maximization) ואת אלגוריתם ויטרבי עבור מודלי מרקוב חבויים, ולאחר מכן את ההתאמות הנדרשות עבור המודל המוכלל שלנו. מימושי האלגוריתמים הללו מוצגים על ידי הפעלתם על מידע סינטטי של רצפים דמויי־מעצמים אשר נוצרו תוך שימוש ביכולת הגנרטיבית של מודל המרקוב החבוי המוכלל. אנו מדמים למודל סביבה מבוקרת על מנת להעריך את ביצועיו בשיערוך פרמטרי המודל והשוואתם לפרמטרים האמיתיים בהם היה שימוש בעת יצירת המידע. לסיום, אנו משתמשים באלגוריתם מיקסום התוחלת על מנת לאמן את המודל שיצרנו על רצפים גנטיים של מעצמים אנושיים, אשר נבחרו לפי המידע האפיגנטי של פרוייקט Roadmap. אנו מדגימים את יכולות המודל על ידי השוואת שיערוכו למידע האפיגנטי של הרצפים, ומביאים ראיה ליכולת המודל לחזות את מיקום המעצמים בגנום ואת הריקמה בהם יופעלו, ללא חשיפה מוקדמת למידע אפיגנטי.