ABSTRACT

Identification of regulatory elements particularly enhancer region plays an important role in comprehending the regulation of gene expression. Current computational enhancer prediction tools are centred at Support Vector Machine (SVM) utilizing sequence content feature—the k-mer. While content feature is shown to be promising, it suffers from several critical weaknesses such as: 1) features associated with enhancer regions are ill-defined and poorly understood. The content feature is unable to represent the complex properties of deoxyribonucleic acid (DNA) sequences; 2) the k-mer feature represents only the global property of DNA sequences but not the localized property; and 3) lack of feature extraction, generation and selection techniques in the algorithm design. This dissertation aims to develop novel feature representations of histone DNA sequences which are associated with enhancer locations. Technical contributions of this study are: 1) complex tree-feature modelling using genetic algorithm (CTreeGA): Automated feature generation framework to capture patterns of interactions among short DNA segments in histone sequences. The interactions of sequence segments are formulated as logical rules and are modelled using a parse-tree. The parse-tree is generated with customized genetic algorithm; 2) k-mer proximity extraction algorithm (KProxEA): Domain dependent feature targeting on spatial relationship among short DNA segments in DNA sequences. The feature models the local distribution of the k-mer in DNA sequences. The locality of feature values is computed using information theoretical formula. Comprehensive evaluations using two histone modification marks, histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac) demonstrate that the proposed features significantly outperformed the sequence content feature. These novel features are able to extract distinguishing patterns significant for classifier learning. Our

findings show that it is necessary to consider complex interaction and locality information of sequence segments in representing DNA sequence feature.

Keywords: Machine Learning, DNA Enhancer Prediction, DNA Feature Representation, Feature Extraction Algorithm.

Ramalan Rangsangan Dna Melalui Teknik Pembelajaran Mesin Dengan Pemodelan Ciri Baharu

ABSTRAK

*Mengenalpasti elemen regulatori terutamanya bahagian rangsangan memainkan peranan yang penting untuk membolehkan pemahaman tentang regulasi mekanisme ekspresi gen. Sistem ramalan komputasi untuk mengenalpasti bahagian rangsangan yang terdapat kini lebih bertumpu pada pendekatan mesin vektor sokongan (SVM) melalui penggunaan ciri urutan kandungan iaitu ciri – k-mer. Walaupun ciri urutan kandungan merupakan ciri yang berkesan, ciri ini mempunyai beberapa kelemahan yang serius seperti: 1) ciri-ciri yang berkaitan dengan bahagian rangsangan tidak jelas dan kurang difahami. Ciri kandungan yang digunakan tidak dapat mewakili sifat-sifat kompleks yang terdapat dalam urutan asid deoksiribonukleik (DNA) 2) ciri k-mer hanya menunjukkan urutan DNA secara meyeluruh dan mengabaikan sifat setempat urutan DNA; dan 3) terdapat kekurangan dari segi pengesktrakan ciri, penjanaan ciri dan teknik pemilihan ciri di dalam formasi algoritma. Tesis ini bermatlamat untuk menghasilkan perwakilan ciri yang baharu untuk urutan histon DNA yang berkaitan dengan lokasi elemen rangsangan. Sumbangan teknikal hasil kajian ini adalah: 1) CTreeGA (pemodelan ciri pokok kompleks menggunakan algoritma genetik): Algoritma penjanaan ciri secara automatik untuk mengekstrak corak interaksi dalam kalangan segmen pendek DNA yang mempunyai urutan histon. Interaksi antara segmen urutan ini diformulasi sebagai peraturan logik ("***logical rules***") dan diwakilkan melalui struktur data pokok. Struktur data pokok ini dihasilkan dengan algoritma genetik yang telah diubahsuai ; 2) KProxE (algoritma ektraksi proksimiti k-mer): Ciri yang bergantung kepada domain yang tertumpu kepada hubungan spatial antara segmen pendek DNA. Ciri ini*

memodelkan taburan local k-mer dalam urutan DNA. Lokaliti nilai-nilai ciri ini diukur melalui komputasi menggunakan formula teoritikal maklumat. Penilaian menyeluruh menggunakan dua tanda modifikasi histon yang berbeza, iaitu H3K4me1 (H3 lysine 4 monomethylation) dan H3K27ac (H3 lysine 27 acetylation) menunjukkan bahawa ciri-ciri yang dicadangkan berjaya mengatasi ciri urutan kandungan . Ciri-ciri baharu ini dapat mengekstrak corak yang berbeza yang penting untuk pembelajaran tentang pengelasan. Penemuan daripada kajian empirikal ini menunjukkan kepentingan mempertimbangkan interaksi kompleks dan informasi lokaliti antara segmen urutan untuk mewakili ciri urutan DNA.

Kata kunci: *Pembelajaran Mesin, Ramalan Rangsangan DNA, Perwakilan Ciri DNA, Algoritma Pengesktrakan Ciri.*