DNA Enhancer Prediction using Machine Learning Techniques with Novel Feature Representation

Fong Pui Kwan

Doctor of Philosophy
2016
DNA Enhancer Prediction using Machine Learning Techniques with Novel Feature Representation

Fong Pui Kwan

A thesis submitted
In fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
2016
DECLARATION

I hereby declare that this thesis entitled DNA Enhancer Prediction using Machine Learning Techniques with Novel Feature Representation is my own original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and list of reference is given.

______________________
FONG PUI KWAN
(13010046)
Department of Cognitive Sciences
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak (UNIMAS)
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr Lee Nung Kion for his continuous guidance and support rendered to me for my Ph.D studies in the Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak. The completion of this thesis would not have been possible without Dr Lee’s motivation, excellent supervision, immense knowledge, dedication and patience throughout the three years of my postgraduate studies. Dr Lee’s guidance has provided me very clear guidelines from the beginning of my studies until the final completion of this thesis with his deep knowledge of computational intelligence and data mining specifically feature extraction, his vast experience in conducting scientific research as well as his professional and positive working attitude which proved to be a strong motivating factor for me to succeed. Dr Lee set a very good example for me in approaching life both as a person and an academician with his endless patience in supervising me and portraying deep passion towards my research. I would like to express my heartfelt thanks and gratitude to Dr Lee’s for his continuous guidance and relentless support in my postgraduate journey as I could not imagine having a better supervisor and mentor for my postgraduate studies.

I would also like to thank my very supportive family members especially my parents, Fong Weng Lun and San Yuen Yee, my husband, Goh Boon Chaw, my grandparents, Goh Cheng Ha and Sun Weng Kong as well as all my family members for their endless support both spiritually and financially in my academic pursuit. I also convey my sincere thanks to my friends, Wong Chin Tieng, Ng Sook Ping, Ho Jong Chan, Sharon Tang, Pang Shek Li and Hoo Soo Ying who continuously gave me words of encouragement and positive vibes throughout my studies. To my laboratory mates Kok WeiYing, Wong Yu Shiong and Ng Sing
Yii, I would like to express my appreciation for all the fruitful discussions and support throughout the three years of pursuing our dreams together.

Last but not least, I would like to express my gratitude and thanks to my scholarship provider, the Ministry of Higher Education, MyBrain (MyPhD) scholarship, everyone who has supported me in my faculty as well as Universiti Malaysia Sarawak lecturers Dr Elena Chai, Associate Professor Dr Fitri Suraya, Associate Professor Dr Teh Chee Siong, Associate Professor Dr Chen Chwen Jen and Ms Oon Yin Bee who have been particularly inspiring and motivating. Special appreciation goes to all my mentors and lecturers who have guided me from my undergraduate study days to the culmination of my postgraduate studies in Universiti Malaysia Sarawak.
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.
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 diformalisasi 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 ekstraksi 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.
TABLE OF CONTENTS

DECLARATION ....................................................................................................................... i

ACKNOWLEDGEMENTS ........................................................................................................ ii

ABSTRACT ............................................................................................................................... iv

ABSTRAK ................................................................................................................................. vi

TABLE OF CONTENTS ........................................................................................................... viii

LIST OF TABLES ..................................................................................................................... xiii

LIST OF FIGURES ................................................................................................................... xvi

LIST OF ABBREVIATIONS ....................................................................................................... xxi

CHAPTER 1: INTRODUCTION ............................................................................................... 1

1.1 Background ...................................................................................................................... 1

1.2 Approaches and Motivations ............................................................................................ 5

1.3 Objectives ......................................................................................................................... 11

1.4 Contributions .................................................................................................................... 11

1.5 Scope of Study ................................................................................................................... 13

CHAPTER 2: BIOLOGICAL BACKGROUND ........................................................................ 15

2.1 Living Cell: Its Characteristics, Properties and Elements .............................................. 15

2.2 Gene Regulation: Process and Interaction of Biological Elements ............................... 18

2.2.1 Trans-regulatory elements and Cis-regulatory elements ........................................... 20

2.3 Epigenetic Modifications .................................................................................................. 22

2.3.1 Histone modifications .................................................................................................. 23

2.4 Relationship between Histone Modifications and Enhancer Regions ......................... 24

2.5 Conclusion ....................................................................................................................... 26
CHAPTER 3: PRELIMINARIES .............................................................................. 28

3.1 Machine Learning Algorithms .................................................................. 28

3.1.1 Introduction ....................................................................................... 28

3.1.2 Supervised Machine Learning Algorithms for Classification ............. 32

3.1.2.1 Support Vector Machine ............................................................... 33

3.1.2.2 Evaluation Metric of Classification Algorithms ............................ 41

3.1.3 Stochastic Learning ............................................................................ 43

3.1.3.1 Genetic Algorithm ........................................................................ 44

3.2 Feature Engineering ................................................................................ 53

3.2.1 Feature construction .......................................................................... 55

3.2.2 DNA sequence representation ............................................................ 55

3.3 Conclusion ............................................................................................. 60

CHAPTER 4: LITERATURE REVIEW ................................................................. 62

4.1 Wet-lab and Experimental Approaches in Identifying Regulatory Elements .... 62

4.2 Early Computational Approaches in Identifying Regulatory Elements .......... 65

4.2.1 Comparative genomics ....................................................................... 65

4.2.2 Motif discovery approaches ................................................................ 69

4.3 Computational Intelligence Approaches for Prediction of Enhancer Regions .... 75

4.3.1 Feature for enhancer region prediction ................................................. 77

4.3.2 Predictive model ................................................................................ 81

4.4 Modelling of Histone Modification Mark Feature ..................................... 83

4.5 Computational Approaches for Feature Construction Framework .............. 86

4.6 Conclusion ............................................................................................. 91
CHAPTER 5: FEATURE CONSTRUCTION: COMPLEX FEATURES

MODELLING USING GENETIC ALGORITHM ........................................... 93

5.1 Introduction ..................................................................................... 93

5.2 Framework Overview .................................................................... 95

5.3 Modelling of Complex Tree-features .......................................... 96

5.4 Complex Tree Generation .............................................................. 102

5.4.1 Terminologies ............................................................................. 103

5.4.2 Initialization of population ......................................................... 104

5.4.3 Encoding complex tree-feature as chromosome ....................... 107

5.4.4 Fitness function .......................................................................... 109

5.4.5 Selection ...................................................................................... 111

5.4.6 Genetic operators ........................................................................ 113

5.4.6.1 Crossover ............................................................................... 113

5.4.6.2 Mutation operator ................................................................. 122

5.4.7 Post processing: Tree-feature selection .................................... 126

5.4.8 Classification ............................................................................... 128

5.5 Experimental Design ..................................................................... 129

5.5.1 Empirical study 1: Comparison study between numbers of top tree-feature for selection of tree-feature set ........................................... 132

5.5.2 Empirical study 2: Comparison study between tree-feature from different CTreeGA generation .................................................. 135

5.5.3 Empirical study 3: Evaluation between proposed tree-feature and baseline conventional feature .................................................. 140
5.5.4 Empirical study 4: Performance of features generated from different DNA dataset………………………………………………………………………………. 146
5.5.5 Analysis of discriminative tree-features ......................................................................................... 147
  5.5.5.1 Distribution of patterns in tree-features .................................................................................. 147
  5.5.5.2 Characteristic of features in selected complex tree-features........................................... 149
5.5.6 Discussion.................................................................................................................. 150
5.6 Summary........................................................................................................................ 153

CHAPTER 6: MODELLING OF K-MER PROXIMITY FEATURE USING INFORMATION THEORY PRINCIPLES ................................................................. 155
6.1 Introduction................................................................................................................................. 155
6.2 System overview.................................................................................................................. 157
6.3 Methods.................................................................................................................................. 158
  6.3.1 Basic concepts and notations............................................................................................... 159
  6.3.2 Modelling of k-mer proximity feature.................................................................................. 160
  6.3.3 Classification .................................................................................................................... 164
6.4 Experimental Design.................................................................................................................. 165
  6.4.1 DNA sequences dataset preparation .................................................................................... 165
  6.4.2 Empirical study 1: Evaluation between k-mer proximity and frequency
      feature ...................................................................................................................................... 167
  6.4.3 Empirical study 2: Comparison between different k-mer proximity
      feature ...................................................................................................................................... 169
  6.4.4 Empirical study 3: Evaluation on the completeness of k-mer proximity
      feature ...................................................................................................................................... 171
CHAPTER 7: DISCUSSION AND FUTURE RESEARCH ........................................ 178

7.1 Discussions ......................................................................................... 179
7.2 Contributions ..................................................................................... 181
7.3 Limitations .......................................................................................... 183
7.3 Future Research .................................................................................. 185

BIBLIOGRAPHY ...................................................................................... 189

APPENDIX A: IMPLEMENTATION OF LIBSVM CLASSIFIER .................. 201

APPENDIX B: DETAIL CROSSOVER PROCEDURE IMPLEMENTED IN
CTREEGA ................................................................................................. 205

APPENDIX C: SIMULATION ON MODELLING OF COMPLEX TREE-
FEATURE: AN ILLUSTRATIVE EXAMPLE ON FITNESS
CALCULATION .......................................................................................... 208

C.0.1 Dataset preparation: positive and negative DNA sequence .............. 208
C.0.2 Feature modelling through CTreeGA .............................................. 209

APPENDIX D: LIST OF PUBLICATIONS ................................................. 211
LIST OF TABLES

Table 3.1  Raw data for classification represented in notations .......................... 32

Table 3.2  Kernel functions to transform data instances from input space to feature
space ................................................................................................................. 38

Table 3.3  Confusion matrix ............................................................................... 41

Table 3.4  Parameters involved in GA ................................................................. 52

Table 3.5  Possible DNA sequence combinations with different length of k-mer ........ 58

Table 4.1  Examples of alignment tools with different properties ......................... 68

Table 4.2  List of common features and its representation for machine learning enhancer
prediction ........................................................................................................... 78

Table 4.3  List of supervised machine learning methods employ in different enhancer
prediction framework ....................................................................................... 82

Table 5.1  Basic building elements in a complex tree representation ..................... 96

Table 5.2  Possible crossover conditions of tree-feature with depth 2, 3 and 4 ........ 114

Table 5.3  Total number of regions formed for each condition of crossover with
tree-feature of dissimilar depth value ............................................................ 115

Table 5.4  Parameters of LibSVM. ...................................................................... 129

Table 5.5  DNA sequences from chromosomes for generating complex tree-feature and
classification ..................................................................................................... 131

Table 5.6  Default parameters for CTree-GA algorithm in generating complex tree-
feature ............................................................................................................... 131

Table 5.7  Comparison of precision rate for prediction using different number of top
tree-features .................................................................................................. 133
Table 5.8  Comparison of recall rate for prediction using different number of top tree-feature. ................................................................. 133

Table 5.9  Comparison of F-measure value for prediction using different number of top tree-features. ................................................................. 133

Table 5.10  CTreeGA parameters used in generating complex tree-features. ................. 137

Table 5.11  Comparisons of precision and recall rates using tree-features generated from different number of generations. ................................................................. 138

Table 5.12  Comparisons of F-measure score for prediction performance using complex tree-features from different maximum number of generations. ................. 138

Table 5.13  Properties of k-mer features. ..................................................................... 141

Table 5.14  Comparisons of precision and recall rates for H3K4me1 prediction using four different top 50% features to represent DNA sequences forming feature vector. ......................................................................................... 143

Table 5.15  Classification results using tree-features generated from chromosome 2........ 147

Table 5.16  Classification results using complex tree-features generated from chromosome 7. ......................................................................................... 147

Table 5.17  Top 5 tree-features from CTreeGA ................................................................ 151

Table 5.18  Examples of tree-features with similar components of short DNA sequences. Similar short DNA segments are highlighted with colours......................... 152

Table 6.1  Performance measure of H3K27ac prediction using two different features...... 168

Table 6.2  Performance measures of H3K27ac prediction using different k-mer proximity ......................................................................................... 169

Table 6.3  Comparisons of prediction accuracy using k-mer proximity and frequency feature .......................................................................................... 170
Table 6.4  Comparison of individual and combined feature generated from 2-mer. ....... 171
Table 6.5  Comparison of individual and combined feature generated from 3-mer. ....... 171
Table 7.1  Summary of contributions. .......................................................... 181
Table A.1  Parameters of LIBSVM. ...............................................................202
Table A.2  LIBSVM parameters setting for classification simulations. .................202
LIST OF FIGURES

Figure 2.1  Double helix DNA structure with nucleotides base pairing. .................................. 16
Figure 2.2  DNA strands’ representation in different direction of double helix. .................. 16
Figure 2.3  Composition of nucleosome and its links. ............................................................... 17
Figure 2.4  Formation of chromosome from coiling and folding of DNA double helix with core histone protein. .............................................................................. 18
Figure 2.5  Simple illustration on orientation of genes in DNA sequence............................ 19
Figure 2.6  Flow diagram for two main steps in gene regulation........................................... 19
Figure 2.7  Overview of cis-acting and trans-acting transcriptional mechanism.................. 19
Figure 2.8  Orientation and location of cis-regulatory elements in the non-coding region of DNA......................................................................................................................... 20
Figure 2.9  Different types of enhancers. ................................................................................... 21
Figure 2.10 Histone modification (Acetylation) changes chromatin structure for binding of transcription factors. ........................................................................................................ 23
Figure 2.11 An example of histone modification: Acetylation on histone tail. ...................... 24
Figure 2.12 Example of spatial relationship between H3K4me1 and H3K27ac with enhancer region. ....................................................................................................................... 26
Figure 3.1  General workflow for machine learning in solving classification problem....... 31
Figure 3.2  Basic principles and concepts of SVM classifier................................................... 36
Figure 3.3  Example of non-linearly separable dataset............................................................ 37
Figure 3.4  Workflow of SVM. .................................................................................................. 39
Figure 3.5  Example of ROC curve. ........................................................................................ 42
Figure 3.6  Different categories of supervised machine learning algorithms for optimization.............................................................................................................................. 44
Figure 3.7  Schematic flowchart of GA for optimization.......................... 46
Figure 3.8  Encoding techniques for possible solutions in optimization search space. ........ 47
Figure 3.9  Summary of Roulette wheel formation and selection of parent using this technique. ........................................................................................................ 49
Figure 3.10  Example of one point binary crossover..................................................... 50
Figure 3.11  Mutation operations............................................................ 51
Figure 3.12  Knowledge discovery process using feature engineering.............................. 54
Figure 3.13  Example of DNA sequence.......................................................... 56
Figure 3.14  Feature construction of 2-mer frequency count........................................... 59
Figure 3.15  DNA sequence represented in normalized k-mer frequency feature......... 60
Figure 4.1  Flow of comparative genomic method to identify functional regions............. 65
Figure 4.2  General workflow for enhancer binding sites prediction using motif discovery programs........................................................................................................ 70
Figure 4.3  Example of consensus derived from binding sites........................................ 71
Figure 5.1  System framework........................................................................ 95
Figure 5.2  General representation of basic feature units............................................. 97
Figure 5.3  General representation of the simplest tree-feature...................................... 97
Figure 5.4  Example on tree-feature comprehension by converting to linear feature representation .................................................................................................................. 99
Figure 5.5  General representation of complex tree-feature with depth 3...................... 100
Figure 5.6  General representation of complex tree-feature with depth 4....................... 100
Figure 5.7  ID numbering for converting complex tree-feature to linear feature representation. ..................................................................................................................... 101
Figure 5.8  Linear representation of complex tree-feature in which pattern-1 and pattern-2 is annotated by $P\text{-}1$ and $P\text{-}2$ respectively while logical operator is annotated by logical operators (LO). ........................................ 102

Figure 5.9  CTree-GA Algorithm .................................................................................. 102

Figure 5.10 Terminologies and concept of tree-feature in CTree-GA algorithm. ......... 103

Figure 5.11 Structure of binary encoding ................................................................. 107

Figure 5.12 An example of how a complex tree is encoded. ...................................... 108

Figure 5.13 Example of targeted DNA sequence with highlighted subsequences indicate matches with k-mer patterns in tree-feature .................................................. 108

Figure 5.14 Procedure in solving the Boolean algebra for complex tree-feature .......... 109

Figure 5.15 Possible points for crossover operation of depth 2 tree-feature represented in (a) parse tree and (b) chromosome................................. 115

Figure 5.16 Possible regions formation represented in parse tree for tree-feature of depth three when crossover with tree-feature of depth two. ......................... 116

Figure 5.17 Crossover between tree-feature with depth 3. .................................. 118

Figure 5.18 Crossover between depth two tree-features ............................................. 119

Figure 5.19 Crossover between depth three tree-features ........................................... 119

Figure 5.20 Crossover between depth four tree-features ............................................. 119

Figure 5.21 Crossover between tree-features with depth 2 and 3 using crossover point 1 and crossover region 2................................................................. 121

Figure 5.22 Crossover between tree-features with depth 2 and 4 using crossover point 2 and crossover region 3................................................................. 121

Figure 5.23 Crossover between tree-features with depth 3 and 4 using crossover point 4 and crossover region 1................................................................. 122
Figure 5.24 Example of mutation operation in basic unit of tree-feature. ........................................ 124
Figure 5.25 Simplified concept on the flow of mutation procedure for complex tree-feature. .......................................................... 125
Figure 5.26 Post processing for top complex tree-feature selection. ........................................ 128
Figure 5.27 Averaged F-measure value of chromosome 2 to 6 in H3K4me1 prediction using different top tree-features. .......................................................... 134
Figure 5.28 Averaged fitness value of tree-features for each generation in GA............. 136
Figure 5.29 F-measure values of H3K4me1 prediction for five chromosomes using three different sets of complex tree-features generated independently with varying number of generations. .......................................................... 139
Figure 5.30 Computation of normalized k-mer frequency feature. This methodology can be generally divided into seven main steps. ........................................ 142
Figure 5.31 Comparisons of F-measure score for prediction of H3K4me1 using different features. ................................................................................. 144
Figure 5.32 Comparisons of auROC values for prediction of H3K4me1 using different features. ................................................................................. 145
Figure 5.33 Composition of two basic patterns in top 500 tree-features. .................. 148
Figure 5.34 Combinatorial composition of pattern-1 and pattern-2 in top 500 selected tree-features. ................................................................................. 150
Figure 6.1 Components involved in machine learning classification for prediction of H3K27ac................................................................. 157
Figure 6.2 Overview of KProxEA System................................................................. 158
Figure 6.3 Concept of spatial relationship represented using proximity score. .......... 159
Figure 6.4  Example of shifting window method to match with 2-mer (GT) in a sequence’s region. .......................................................... 161

Figure 6.5  Matrix representation to store frequency of 2-mer for each region of DNA sequence. .......................................................... 161

Figure 6.6  Variables involved in computing relative frequency. ........................................... 162

Figure 6.7  Example of entropy value for DNA sequences with different k-mer distribution. .................................................................. 163

Figure 6.8  An example on computing proximity feature score for k-mer TA in two sequences. ........................................................................ 164

Figure 6.9  Example on extraction of DNA sequence based on given coordinates. ........ 166

Figure 6.10 ROC curves for prediction using 2-mer proximity and frequency feature. ...... 168

Figure 6.11 Comparisons of F-measure score for k-mer proximity and frequency feature. ...................................................................... 171

Figure 6.12 Comparisons of F-measure score using different k-mer features for H3K4me1 prediction. ................................................... 173

Figure 7.1  K-mers from different region located closely with each other. ..................... 184
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>AlignACE</td>
<td>Aligns Nucleic Acid Conserved Elements</td>
</tr>
<tr>
<td>auROC</td>
<td>Area under Receiver Operating Curve</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>Chromatin immunoprecipitation microarray</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation sequencing</td>
</tr>
<tr>
<td>CpG</td>
<td>C–phosphate–G</td>
</tr>
<tr>
<td>C-SVC</td>
<td>C-Support Vector Classifier</td>
</tr>
<tr>
<td>CTreeGA</td>
<td>Complex tree-feature modelling using genetic algorithm</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DHS</td>
<td>DNase I hypersensitivity sites</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopaedia of DNA elements</td>
</tr>
<tr>
<td>epsilon-SVR</td>
<td>epsilon-Support Vector regression</td>
</tr>
<tr>
<td>EP300</td>
<td>E1A binding protein p300</td>
</tr>
<tr>
<td>FESP</td>
<td>Frequent Emerging Sequence Patterns</td>
</tr>
<tr>
<td>FG-EA</td>
<td>Feature Generation Evolutionary Algorithm</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>IUPAC</td>
<td>International Union of Pure and Applied Chemistry</td>
</tr>
<tr>
<td>H3Ac</td>
<td>H3 acetylation</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>H3 lysine 4 monomethylation</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>H3 lysine 4 trimethylation</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>H3 lysine 27 acetylation</td>
</tr>
<tr>
<td>KProxEA</td>
<td>K-mer proximity extraction algorithm</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Shrinkage and Selection Operator</td>
</tr>
<tr>
<td>LO</td>
<td>Logical Operator</td>
</tr>
<tr>
<td>MDscan</td>
<td>Motif Discovery scan</td>
</tr>
<tr>
<td>MEME</td>
<td>Multiple EM for Motif Elicitation</td>
</tr>
<tr>
<td>MNIST</td>
<td>Mixed National Institute of Standards and Technology</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger Ribonucleic acid</td>
</tr>
<tr>
<td>NP-hard</td>
<td>Non-deterministic Polynomial-time hard</td>
</tr>
<tr>
<td>nu-SVC</td>
<td>nu- Support Vector Classifier</td>
</tr>
<tr>
<td>nu-SVR</td>
<td>nu-Support Vector Regression</td>
</tr>
<tr>
<td>PFM</td>
<td>Position Frequency matric</td>
</tr>
<tr>
<td>PhastCon</td>
<td>Phylogenetic Analysis with Time model Conservation</td>
</tr>
<tr>
<td>PHP</td>
<td>Hypertext Preprocessor</td>
</tr>
<tr>
<td>PWM</td>
<td>Position Weight Matric</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Curve</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription factor binding site</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California, Santa Cruz</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>

xxii