Motif based computational identification of protein subcellular localisation

Mutlu Doḡruel

Wellcome Trust Sanger Institute

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

January 2008

To my deeply missed father,

Sami Han Doḡruel.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Acknowledgements

I would like to express my gratitude to Dr. Tim Hubbard for being a great supervisor. His support, invaluable suggestions and comments during our weekly meetings are much appreciated. I am also deeply indebted to my PhD committee members Dr. Ewan Birney and Dr. Ian Dunham for their time and advice. I would also like to thank my examiners Prof. Søren Brunak and Dr. Nick Goldman for reading the thesis thoroughly and for their useful feedback on the preliminary version of the thesis.

My sincere thanks go to Dr. Thomas Down for his useful suggestions and also patience during the discussions we made (usually accompanied with a cup of coffee) on Biojava, NestedMICA, Eponine and many other topics... Many thanks also to the other members of Tim's research group at the Sanger Institute: Dr. Andreas Prlić, and will-be Drs. Markus Brosch, Matias Piipari, and Jenny Mattison for making my PhD days more enjoyable, and being excellent friends. I think Markus did a great job by encouraging me to join the Cambridge University gliding club. Being in the air over the weekends in the summer was a good way to relax and get ready for the productive work days;)

Thanks to Tim, Matias, Andreas and Jenny again for critically reading the preliminary versions of the chapters. I also thank Antony Quinn from EBI for his suggestions about the thesis.

I also would like to thank the Sanger Institute's web, systems and library staff for their help on various occasions.

I am grateful to the Wellcome Trust for financing me during my 4 years in Cambridge.

Finally, I am deeply and forever indebted to my parents, sisters and brother for their love, support and encouragement throughout my entire life: Sizi çok seviyorum!

Mutlu Doḡruel, January 2008, Cambridge, UK.

Abstract

Discovering overrepresented patterns in amino acid sequences is an important step in protein functional annotation which includes the identification of subcellular localisation. I adapted and extended NestedMICA, an *ab initio* protein motif finder originally developed for finding transcription binding site motifs, to find short protein signals, and compared its performance with another popular protein motif finder, MEME.

In order to assess NestedMICA as a protein motif finder, I have tested it on synthetic datasets produced by spiking instances of known motifs from protein databases into a randomly selected set of protein sequences. Apart from the artificially implanted motifs, NestedMICA also successfully recovered subcellular localisation signals from biologically-authentic test sets. NestedMICA found most of the short test protein motifs spiked into a test set of sequences at different frequencies. In all the assessment experiments, its overall motif discovery performance was better than that of MEME.

As a practical application of NestedMICA, I developed a novel Support Vector Machines based protein subcellular classification tool, Lokum, for eukaryotic protein subcellular localisation prediction, covering all major localisation classes for animal, fungal and plant sequences. It uses targeting and retention signal motifs reported by NestedMICA, and other protein features including transmembrane topologies and amino acid composition. Additionally, in Lokum I use bipartite nuclear localisation signals obtained by adding protein support to Eponine, a tool originally developed for transcription start site modeling. Lokum does not use sequence similarity, or any other *a priori* knowledge such as known nuclear localisation signals by searching databases.

I compared proteins targeted into the nuclei and nucleoli in terms of the features used in Lokum, and also their predicted disorder regions. I demonstrate that it is possible to computationally distinguish these two sub-nuclear protein categories.

Finally, as an alternative to the transmembrane topology predictor TMHMM that is used in Lokum, I designed and tested a new prototype program that is based on hidden Markov models (HMM). The HMM has been trained by a novel, nested sampling based transition probability optimisation procedure.

Contents

1 Introduction **1**

1.1 General introduction 1

1.1.1 Previous work on subcellular localisation prediction 6

1.2 Sequence identity thresholds 12

1.3 Computational methodologies 15

1.3.1 HMMs 15

1.3.2 The general idea behind motif finding 18

1.3.3 Inference by Nested Sampling 20

1.3.4 Support vector machines 22

2 NestedMICA as an *ab initio* protein motif finder **25**

2.1 Background 25

2.2 Materials and methods 28

2.2.1 NestedMICA 28

2.2.1.1 The NestedMICA sequence model 29

2.2.1.2 Implementation of NestedMICA 31

2.2.2 Adding protein support to NestedMICA 33

2.2.3 Program output and sequence logos 34

2.2.4 Background model training 35

2.2.5 Testing NestedMICA's performance 36

2.3 Results and discussions 41

2.3.1 Protein sequence background model 41

2.3.2 Performance vs. motif abundance 42

2.3.3 Performance with multiple motifs 48

2.3.4 Performance vs. protein length 53

2.3.5 "Null test" and significance of motifs 55

2.3.6 Testing non-*ab initio* motif finders 58

2.4 Conclusions 59

2.5 Availability and requirements of NestedMICA 61

3 Lokum: *ab initio* protein subcellular localisation prediction for eukaryotes by using mono and bipartite motifs, transmembrane protein topologies, and amino acid composition 63

3.1 Introduction 63

3.1.1 Features used in Lokum 64

3.1.2 Predicted classes 69

3.2 Materials and methods 69

3.2.1 Localisation motif discovery with NestedMICA 69

3.2.2 Motif selection 71

3.2.3 Using Eponine with NestedMICA for multi-component motif discovery 73

3.2.4 Using amino acid composition 77

3.2.5 Using transmembrane topology predictions 78

3.2.6 Training and testing of SVM 79

3.2.7 Evaluation of Lokum predictions 80

3.3 Results 80

3.3.1 Discovered monopartite motifs 82

3.3.1.1 Contribution of N-linked glycosylation signal . . 85

3.3.1.2 Alternative ER retrieval 86

3.3.1.3 Scanning motifs in certain positions 89

3.3.1.4 Scoring multiple instances of motifs 89

3.3.2 Bipartite motif models 91

3.3.2.1 Bipartite NLS 91

3.3.2.2 Bipartite PTS2 92

3.3.3 Golgi N-terminal transmembrane topology prediction statistics help in localisation prediction 94

3.3.4 Effect of amino acid composition 95

3.3.5 Lokum's performance 97

3.3.6 Contributions of different features 98

3.3.7 Contribution of disordered region predictions 102

3.4 Discussions 103

3.5 Availability 105

4 Discriminating nucleolar proteins from nuclear proteins: is it possible? 107

4.1 Introduction 107

4.1.1 Disordered protein regions 111

4.1.2 Protein disorder region prediction 112

4.2 Materials and methods 113

4.2.1 Datasets 113

4.2.2 Training background models for nucleolar and nuclear datasets 116

4.2.3 Running RONN 118

4.2.4 Training the SVM 118

4.3 Results 120

4.4 Discussions and conclusions 131

5 Predicting protein transmembrane topology and signal peptides: An HMM approach with a new parameter optimisation strategy 134

5.1 Introduction 134

5.1.1 The aim of this study 134

5.1.2 Transmembrane topology and signal peptide prediction . . 136

5.2 Materials and methods 139

5.2.1 Architecture of the HMM 139

5.2.1.1 Representing helix caps in the HMM 142

5.2.1.2 Duration HMM states 147

5.2.2 Datasets and training of the model 148

5.2.3 Transition probability optimisation: a new approach 150

5.3 Results 159

5.3.1 Signal peptides at the DNA level 161

5.4 Discussions 163

6 Conclusions 165

A Motifs discovered by NestedMICA 170

B Amino acid composition rates in different localisations 180

C Sequence IDs of nuclear and nucleolar proteins filtered from the LOCATE database 191

C.1 Proteins in nucleoli 191

C.2 Proteins in nuclei 194

D Kullback-Leibler divergence for transmembrane helix cap positions in terms of amino acid composition 197

References 233

List of Figures

1.1 The “thesis graph” 4

1.2 Probability of staying in the same state in a minimum duration capable HMM state 17

1.3 Profile HMMs 18

2.1 Likelihood curve for different number of mosaic classes 37

2.2 Motifs recovered by NestedMICA and MEME in the single-motif spiking tests, for motif set1 43

2.3 Motifs recovered by NestedMICA and MEME in the single-motif spiking tests, for motif set2 44

2.4 Motifs recovered by NestedMICA and MEME in the single-motif spiking tests, for motif set3 45

2.5 Inserting more than one different motif into the sequences. 51

2.6 Motif recovery performance against sequence length 55

2.7 NestedMICA’s “null motifs” 57

2.8 A snapshot showing the regular expressions reported by the Dilimot web service 60

3.1 Eponine TSS model 74

3.2 Eponine TTS model 75

3.3 SVM kernel parameter optimisation 81

3.4 Some of the protein localisation related signals as recovered by NestedMICA. 83

3.5 Unannotated or less known motifs found by NestedMICA. 84

3.6 Manually built motifs that are used in the ChloroP predictor. . . 85

3.7 Contribution of the N-linked glycosylation motif in localisation classification 87

3.8 The two C-terminal ER retention motifs reported 88

3.9 A ROC curve showing the effect of scanning sequences with PWMs in certain segments only 90

3.10 Schematic representation of the Eponine bipartite NLS model . . 92

3.11 The expected number of amino acids in the first 60 N-terminal residues to be part of a transmembrane region. 96

3.12 Individual contributions of features used in the SVM 101

3.13 Lokum prediction service hosted by the Wellcome Trust Sanger Institute 106

4.1 Nuclear pore 109

4.2 NestedMICA motifs discovered from nuclear and nucleolar datasets 117

4.3 A protein disorderded regions plot based on RONN predictions . . 119

4.4 A selection of the protein motifs recovered by NestedMICA from a set of nuclear and a set of nucleolar proteins, using a cytoplasmic background 121

4.5 Score histograms for N- and C-terminal nucleolar motifs 122
4.6 Distributions of amino acids predicted to be within TM helices in nuclear and nucleolar proteins 124
4.7 Differences between nucleolar and nuclear proteins in terms of their amino acid compositions 125
4.8 Score distribution of a core nucleolar motif within nuclear and nucleolar proteins 127
4.9 A larger number of nucleolar localisation signal hits can be found in disordered regions 129
4.10 Generally a larger number of NLS signal hits can be found in disordered regions of nucleolar proteins compared to nuclear sequences 130

5.1 The architecture of the developed transmembrane predictor 141
5.2 Total likelihood function monotonically increases in nested sampling 153
5.3 Transition probability set having the least likelihood is replaced by new better one 154
5.4 It becomes more difficult over time to find "acceptable" states . . 155
5.5 Signal peptide motif at the RNA level 162

A.1 "Nuclear motifs" discovered by NestedMICA 171
A.2 "Plasma membrane motifs" discovered by NestedMICA 172
A.3 "Cytoplasmic motifs" discovered by NestedMICA 173
A.4 "Mitochondrial motifs" discovered by NestedMICA 174
A.5 "Endoplasmic reticulum motifs" discovered by NestedMICA . . . 175
A.6 "Golgi motifs" discovered by NestedMICA 176

A.7 “Extracellular motifs” discovered by NestedMICA 177

A.8 “Lysosome motifs” discovered by NestedMICA 178

A.9 “Peroxisomal motifs” discovered by NestedMICA 178

A.10 “Vacuolar motifs” discovered by NestedMICA 179

B.1 Amino acid composition for Alanine (ALA / A) 181

B.2 Amino acid composition for Arginine (ARG / R) 181

B.3 Amino acid composition for Asparagine (ASN / N) 182

B.4 Amino acid composition for Aspartic Acid (ASP / D) 182

B.5 Amino acid composition for Cysteine (CYS / C) 183

B.6 Amino acid composition for Glutamine (GLN / Q) 183

B.7 Amino acid composition for Glutamic Acid (GLU / E) 184

B.8 Amino acid composition for Glycine (GLY / G) 184

B.9 Amino acid composition for Histidine (HIS / H) 185

B.10 Amino acid composition for Isoleucine (ILE / I) 185

B.11 Amino acid composition for Leucine (LEU / L) 186

B.12 Amino acid composition for Lysine (LYS / K) 186

B.13 Amino acid composition for Methionine (MET / M) 187

B.14 Amino acid composition for Phenylalanine (PHE / F) 187

B.15 Amino acid composition for Proline (PRO / P) 188

B.16 Amino acid composition for Serine (SER / S) 188

B.17 Amino acid composition for Thereonine (THR / T) 189

B.18 Amino acid composition for Tryptophan (TRP / W) 189

B.19 Amino acid composition for Tyrosine (TYR / Y) 190

B.20 Amino acid composition for Valine (VAL / V) 190

List of Tables

1.1 A list of some popular eukaryotic localisation predictors 7

1.2 Maximum mutual sequence identity rates allowed in the different predictors 13

1.3 Several allowed maximum pairwise sequence identity rates versus the number of vacuolar sequences 14

2.1 Motif recovery performance for NestedMICA and MEME for individual test sets. 46

2.2 Total motif recovery performance summary for NestedMICA and MEME. 47

2.3 Sensitivity and specificity values for motifs of Set 1, reported by NestedMICA and MEME in the single-motif spiking tests 48

2.4 Sensitivity and specificity values for motifs of Set 2, reported by NestedMICA and MEME in the single-motif spiking tests 49

2.5 Sensitivity and specificity values for motifs of Set 3, reported by NestedMICA and MEME in the single-motif spiking tests 50

2.6 Performance summary for NestedMICA in the multiple motif spiking tests. 52

2.7 Performance summary for MEME in the multiple motif spiking tests. 53

2.8 Sensitivity and specificity values for motifs reported by Nested-MICA in the multiple-motif spiking tests 53

2.9 Sensitivity and specificity values for motifs reported by MEME in the multiple-motif spiking tests 54

3.1 Sequences used in the motif discovery phase 71

3.2 Protein background parameters for datasets used in localisation motif discovery 72

3.3 Alternative amino acid groupings used in composition calculation 78

3.4 Prediction performance summary for Lokum. 99

3.5 MCCs and correct prediction rates for Lokum, MultiLoc and PSORT 100

4.1 Performance measures calculated from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins of CASP 6 114

5.1 Amino acid emission probabilities in the transmembrane helix cytoplasmic side cap. 142

5.2 Amino acid emission probabilities in the transmembrane helix non-cytoplasmic side cap. 143

5.3 KL deviations of cytoplasmic side helix termini from the noncytoplasmic side helix termini in relative amino acid abundance rates. 144

5.4 Minimum allowed emission state occupancy numbers for the transmembrane topology predicting HMM 149

5.5 Sequences used in the training of Phobius and of the program developed 149

5.6 Prediction performance summary for “TM-and-SP”, “TM-only” and “non-SP, non-TM” proteins. 160

5.7 Correct prediction rates for SP-only proteins. 160

C.1 Nucleolar proteins 194

C.2 Nuclear proteins 196

D.1 KL deviations of cytoplasmic side helix cap position 1 from the non-cytoplasmic side helix cap positions in relative amino acid abundance rates. 198

D.2 KL deviations of cytoplasmic side helix cap position 2 from the non-cytoplasmic side helix cap positions in relative amino acid abundance rates. 199

D.3 KL deviations of cytoplasmic side helix cap position 3 from the non-cytoplasmic side helix cap positions in relative amino acid abundance rates. 200

D.4 KL deviations of cytoplasmic side helix cap position 4 from the non-cytoplasmic side helix cap positions in relative amino acid abundance rates. 201

D.5 KL deviations of cytoplasmic side helix cap position 5 from the non-cytoplasmic side helix cap positions in relative amino acid abundance rates. 202

D.6 KL deviations of cytoplasmic side helix cap positions. 203

D.7 KL deviations of cytoplasmic side helix cap positions. 204

D.8 KL deviations of non-cytoplasmic side helix cap positions. 205

D.9 KL deviations of non-cytoplasmic side helix cap positions. 206