Machine-Learning based sequence analysis, bioinformatics & nanopore transduction detection

Stephen Winters-Hilt

University of New Orleans
New Orleans, Louisiana
&
Meta Logos Inc.
New Orleans, Louisiana

ISBN 978-1-257-64525-1

Lulu Publication

Preface

This is intended to be a simple and accessible book on machine learning methods and their application in computational genomics and nanopore transduction detection. This book has arisen from the past eight years of teaching one-semester courses on various machine-learning, cheminformatics, and bioinformatics topics. Possible uses of this textbook in one-semester courses are as follows:

(1) Introductory Bioinformatics – an undergraduates level course covered in Ch.s 1-4, if not delving into the HMM derivations in Ch. 3 extensively. A more advanced one-semester Bioinformatics course, for senior undergraduates and graduate students, can be based on Ch.s 1-4 if the HMM derivations are pursued in detail.

(2) Introductory Machine Learning – a one-semester course on HMMs, SVMs, and their pattern recognition and structure discovery applications, at the senior undergraduate or graduate student level, can be based on Ch.s 1-3, 4.D, 5, 8, and 9.

(3) Introductory Stochastic Signal Analysis – a one-semester course on HMMs and SVMs and their signal processing applications, at the senior undergraduate or graduate student level, can be based on Ch.s 1-3, 5, 8, and 9.

(4) Introductory Nanopore Transduction Cheminformatics – a one-semester course on Nanopore Transduction Detection and related Channel Current Cheminformatics, at the senior undergraduate or graduate student level, can be based on Ch.s 1-3, and 5-7.

(5) Advanced HMMs -- a one-semester course on Hiddeen Markov Models, at graduate student level, can be based on Ch.s 1-3, 8, and 9.

(6) Advanced SVMs -- a one-semester course on Support Vector Machines, at graduate student level, can be based on Ch.s 1,2,5,9.

This book partly draws on material that the author that has published in open source journals, including material from the EURASIP Journal for Advanced Signal Processing [1,2], and from BMC Bioinformatics publications [3-18].

Stephen Winters-Hilt

New Orleans, March 2011

Acknowledgements

I'd like to thank the student collaborators and lab technicians I've worked with, and co-authored papers with, over the past eight years. In particular, I would like to thank lab technicians Amanda Alba, Amanda Davis, Andrew Duda, Iftekhar Amin, and, especially, Eric Morales, for help performing the nanopore experiments, and I'd like to thank a host of University of New Orleans and Tulane undergraduate, graduate, and postdoctoral students for help with the nanopore experiments and the channel current cheminformatics analysis. In recent efforts this includes Evenie Horton, Jorge Chao, Joshua Morrison, and in prior collaborations this includes:

Anand Prabhakaran (UNO; MS in CS, Fall 2005). Thesis title: Power Signal Analysis of Channel Current signal using HMM-EM and time domain FSA;

Alex Ortiz (Tulane; MS in Biomedical Engineering, Spring 2006). Thesis title: DNA Binding Characterization of Pseudo Aptamers using Nanopore Technology;

Raja Iqbal (Tulane; PhD in Computer Science, Spring 2006). Thesis title: Robust Learning Algorithms: Applications in Data Mining, Computer Vision and Bioinformatics;

Srikanth Sendamangalam (UNO; MS in CS, Summer 2006). Thesis title: Nanopore Detector Feedback Control Using Cheminformatics Methods Integrated with LabView/LabWindows Tools;

Charlie McChesney (UNO; MS in CS, Summer 2006). Thesis title: SVM-based Clustering;

Matthew Landry (UNO; MS in CS, Spring 2007). Thesis title: Analysis of Nanopore Detector Measurements using Machine Learning Methods, with application to single-molecule kinetics;

Molly Oehmichem (Tulane; BS in Biomedical Engineering, Spring 2008). Thesis title: Distinction of Single Nucleotides for the purpose of DNA sequencing using a nanopore-based detector;

Kenneth Armond Jr. (UNO; MS in CS, Summer 2008). Thesis title: Distributed Support Vector Machine Learning;

Sepehr Merat (UNO; MS in CS, Summer 2008). Thesis title: Clustering via supervised support vector machines;

Daming Lu (UNO; MS in CS, Summer 2009). Thesis title: Motif Finding;

Hang Zhang (UNO; MS in CS, Summer 2009). Thesis title: Distributed Support Vector Machines with Graphical Processing Units;

Carl Baribault (UNO; PhD in CS, Fall 2009). Thesis title: Meta-state generalized HMMs for eukaryotic gene structure identification;

Zuliang Jiang (UNO; PhD in CS, Spring 2010). Thesis title: Binned HMM with duration: variations and applications;

Alexander Churbanov (Postdoc, 2006-2008); and

Alexander Stoyanov (Postdoc, 2006-2008).

I'd like to thank the University of New Orleans, NIH, NSF, NASA, and the Louisiana Board of Regents for research support. The author would also like to thank META LOGOS Inc., for research support and a research license.

META LOGOS INC. was co-founded by the author in 2009, when it obtained exclusive license to the nanopore transduction detector (NTD) and machine-learning based signal processing intellectual property. The author would like to thank Robert Adelman (CEO META LOGOS, Inc.), Andrew Peck (CEO PxBioSciences), and Mike Lewis (Professor, University of Missouri-Columbia), for discussion exploring the potential impact of the NTD approach. Prior to incorporation, Meta Logos was a sole-proprietorship, Meta Logos Systems, specialized in machine learning based signal processing, and was founded by the author in 1997 in Santa Cruz, CA.

Stephen Winters-Hilt

Contents