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

Copyright © 2011 by Stephen Winters-Hilt. All rights reserved.

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 Hidden 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
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:

- 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;
- Charlie McChesney (UNO; MS in CS, Summer 2006). Thesis title: SVM-based Clustering;
Molly Oehmicem (Tulane; BS in Biomedical Engineering, Spring 2008). Thesis title: Distinction of Single Nucleotides for the purpose of DNA sequencing using a nanopore-based detector;


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

List of Figures xix

List of Tables xxv

1 Introduction 1

1.A Stochastic Sequential Analysis via Hidden Markov Models 2
1.B Support Vector Machines for Classification and Clustering 4
1.C Nanopore transduction detection 5

2 Ad hoc signal recognition using Information Theory 7

2.A Information Theory Methods 7
  2.A.1 The Calculus of Conditional Probabilities 8
  2.A.2 Frequentist vs. Bayesian Statistics 9
  2.A.3 Shannon entropy: the Khinchin derivation 9
  2.A.4 Relative Entropy Uniqueness 10
  2.A.5 Mutual Information 10
  2.A.6 Information measures 10
  2.A.7 Significant Distributions that are not Geometric 10
  2.A.8 Significant Series that are Martingale 12
  2.A.9 Markov Chains 12

2.B Standard Electrical Engineering signal analysis Tools 14
2.B.1 Nyquist Sampling Theorem 15
2.B.2 Fourier Transforms, and other Transforms 15
2.B.3 Power spectral density (PSD) 15
2.B.4 Cross-PSD 16
2.B.5 AM/FM/PM Communications Protocol 16
2.B.6 Phase-locked loop (PLL) Protocol 17

2.C Ad Hoc Methods 18
2.C.1 Channel Current Cheminformatics (CCC) Protocol 18
2.C.2 Finite State Automata (FSAs) with holistic tuning 19
2.C.3 tFSA spike detector 29
2.C.4 Mutual Information linkage identification 31
2.C.5 *Ab initio* learning with holistic and bootstrap learning 32
2.C.6 Comparative topological structure identification 34

2.D Problems 37

3 Analysis of Stochastic Data using Hidden Markov models 41

3.A Hidden Markov model (HMM) Background 42
3.A.1 When to use a Hidden Markov Model (HMM)? 47
3.A.2 Weaknesses of the standard HMM 48

3.B Hidden Markov models and HMM-based feature extraction 52
3.B.1 Viterbi Path 53
3.B.2 Forward and Backward Probabilities 54
3.B.3 HMM: Maximum Likelihood discrimination 55
3.B.4 Expectation/Maximization 55
3.B.5 Emission and Transition Expectations with Rescaling 56
3.B.6 pMM/SVM 57
3.B.7 Feature Extraction via EVA projection 58
3.B.8 Feature Extraction via Data Absorption 59
3.B.9 Modified AdaBoost for Feature selection and fusion 60
3.B.10 HMM/Viterbi Code examples (in C) 64

3.C Linear HMM 71
3.D The Meta-HMM – a clique-generalized HMM 73
3.E Hidden Semi-Markov model and HMM-with-duration 81
3.F HMM with binned duration 87
3.G Distributed HMM with possible GPU speedup 91
3.H Problems 93

4 Bioinformatics 95
4.A. Development of chemical replicators & info. structures  
4.A.1 Formation of the Pre-Life Physical Environment  
4.A.3 Role of Viruses and other ‘selfish’ genomic elements  
4.A.4 Role of Artefact  
4.A.5 Encapsulated interactions and information structures  

4.B. Information encoding molecules and structures (genomes)  
4.B.1 DNA  
4.B.2 mRNA  
4.B.3 Protein  
4.B.4 Genomes  
4.B.4.1 Virus Genomes  
4.B.4.2 Prokaryotic Genomes  
4.B.4.3 Eukaryotic Genomes  

4.C. Bioinformatics Methods  
4.C.1 Electrophoresis and GELs  
4.C.2 PCR  
4.C.3 DNA Sequencing  
4.C.4 Expression analysis: DNA microarrays & RNA-seq  
4.C.5 BLAST: sequence alignment  
4.C.6 Phage Typing & Metagenomic Testing  

4.D. Computational Genomics  
4.D.1 Meta-HMM for eukaryotic genome analysis  
4.D.1.1 Primer on Genomic Data – C. elegans specifics  
4.D.1.2 The meta-HMM model for genomic analysis  
4.D.1.3 HMM states for gene-structure identification  
4.D.1.4 Measures of Predictive Performance  
4.D.1.5 Meta-HMM Results  
4.D.2 HMMBD+pde+zde Eukaryote  
4.D.2.1 ZoneDependent Emission (ZDE) modeling  
4.D.2.2 HMMBD+pde+zde analysis on C. elegans  
4.D.3 Preliminary Alt-splice model  
4.D.3.1 Two-track alt-splice gene finder model  
4.D.3.2 Statistical support for Alt-splice two-track model  

4.E Problems  

5 Classification & Clustering using Support Vector Machines
5.A Decision Boundary and SRM Construction using Lagrangian 158
5.A.1. The theory of classification 162
5.A.2. Kernel modeling and other Tuning 164
5.A.3 Kernel construction using polarization 165
5.A.4 Support Vector Machine Lagrangian formulation 168

5.B SVM Kernels 173
5.B.1 The ‘stable’ kernels 176
5.B.2 Entropic and Gaussian Kernels 179

5.C SVM optimization using SMO and alpha-selection heuristics 180
5.C.1 Sequential Minimal Optimization (SMO) 180
5.C.2 Code Samples 184
5.C.3 Adaptive Feature Extraction/Discrimination 192
5.C.4 Robust SVM performance in the presence of noise 193

5.D Multiclass SVM 193
5.D.1 SVM-External Multiclass 194
5.D.2 SVM-Internal Multiclass 195
5.D.3 SVM Speedup via differentiating BSVs and SVs 197

5.E. SVM Chunking and Tuning 199
5.E.1 Tuning 200
5.E.2 SVR Method 200
5.E.3 Chunking Protocols 201
5.E.4 Chunking Pathologies 202
5.E.5 SVM Distributed processing with GPU/CPU 203

5.F Data-rejection heuristics 204
5.F.1 Data Rejection Tuning 204
5.F.2 Marginal Drop with SVM-Internal 206

5.G SVM Clustering 206
5.G.1 SVM ‘Internal’ Clustering 208
5.G.2 SVM-External Clustering 211
5.G.3 SVM-External Clustering – Algorithmic Variants 224
5.G.4 Binary classifier & clustering scoring conventions 229

5.H Problems 231

6 Single-molecule Biophysics and Nanopore Detection 233
6.A Protein Channel Electrochemistry and Biophysics

6.A.1 Thermodynamics in Biophysics
   6.A.1.1 Equilibrium
   6.A.1.2 Non-equilibrium
   6.A.1.3 Fluid Flow
   6.A.1.4 Absolute Reaction Rate

6.A.2 Simple ions in solution
   6.A.2.1 Water Clustering
   6.A.2.2 Hydration Radius
   6.A.2.3 Debye Radius

6.A.3 DNA and polymer ions in solution
   6.A.3.1 DNA Structure from crystallography and NMR
   6.A.3.2 Discerning Structure of Duplex Ends

6.A.4 Membranes and channels
   6.A.4.1 Nanopores in Lipid Bilayers
   6.A.4.2 The highly stable, α-hemolysin protein channel
   6.A.4.3 Membrane Environment in Biosensing

6.A.5 The Coulter Counter

6.A.6 Partitioning and Translocation in Channels
   6.A.6.1 The Free Energy Barrier
   6.A.6.2 ssDNA partitioning/translocation in α-HL
   6.A.6.3 Temperature effects

6.A.7 Forces acting on polymers in a nanopore

6.A.8 Engineered and Synthetic Channels

6.B Nanopore Detector Biophysics

6.B.1 Protein Channel Electrochemistry Environment
   6.B.1.1 Standard and physiological buffer conditions
   6.B.1.2 α-Hemolysin stability – use of chaotropes
   6.B.1.3 The Nanopore Detector Voltage Clamp Circuit
   6.B.1.4 Nanopore Detector Electronic Noise Sources
   6.B.1.5 Controlling α-HL Noise via Choice of Aperture

6.B.2 Nanopore Detector Blockade Sensing
   6.B.2.1 The α-HL nanopore blockade detector
   6.B.2.2 Nanopore biosensor single-signal saturation
   6.B.2.3 Nanopore Detector Membrane Stability
   6.B.2.4 Bandwidth limitations
   6.B.2.5 Sticking problem and use of Excitations
   6.B.2.6 Other Single Molecule Methods

6.B.3 Mechanism of modulatory channel blockades
6.B.3.1 The 9bp hairpin blockade mechanism
6.B.3.2 Conformational Kinetics on Model Biomolecules

6.C Summary of translocation time ND biosensing methods

6.D Things to ‘contact’ with the channel other than ssDNA
6.D.1 Aptamers
6.D.2 Bifunctional Immunoglobulins

6.E Channel Current Cheminformatics (CCC) Methods

6.F Problems

7 The Nanopore Transduction Detector ‘Nanoscope’

7.A NTD Background
7.A.1 Nanopore Transduction Detection
7.A.2 Bifunctional NTD aptamers
7.A.3 Potential Impact

7.B NTD Platform and Operation
7.B.1 NTD Operational Protocol
7.B.2 Driven modulations
7.B.3 Driven modulations with multichannel

7.C NTD Biosensing Methods & Proof of Concept Results
7.C.1 Model system based on streptavidin and biotin
7.C.2 Model system based on DNA annealing
7.C.2.1 (preliminary) linear DNA annealing test
7.C.2.2 (preliminary) ‘Y’ DNA annealing test
7.C.3 Pathogen Detection
7.C.4 SNP Detection
7.C.5 Aptamer-based Detection
7.C.6 Antibody-based Detection
7.C.6.1 Small target Antibody-based detection
7.C.6.2 Large target Antibody-based detection

7.D NTD Assaying Methods & Proof of Concept Results
7.D.1 DNA enzyme analysis: Integrase
7.D.2 Single-molecule serial assaying
7.D.2.1 Glycoprotein assayer
7.D.2.2 Antibody Assay: A window into Ab function
7.D.2.3 Multicomponent Molecular Analyzer 314
7.D.3 Molecular capture & TERISA 315
7.D.4 NTD-Gel 317
7.D.5 Nanopore Processing Unit (NPU) 317
7.D.6 Immunological Screening using CCC 318
7.D.7 Assays of cytosolic antigen delivery complexes 319

7.E Specific Application Areas 319
7.E.1 Nanopore Transduction Platform and Carrier Signal 319
7.E.2 Model systems for NTD Assaying: Aptamer-TBP 320
7.E.3 Deciphering the Transcriptome & Drug Discovery 322
7.E.4 DNA Sequencing 323
7.E.4.1 Single-molecule, processive 323
7.E.4.2 NTD/Sanger DNA Sequencing 327

7.F List of NTD Proof-of-concept Experiments 328

7.G Problems 333

8 Stochastic sequential analysis, classification, & clustering 335
8.A Stochastic Sequential Analysis (SSA) 335
8.A.1 CCC implementation of the SSA protocol 336
8.A.2 NTD: a binary stochastic ‘phase’ modulation 336

8.B The SSA Protocol 339

8.C SCW for boosting and secure, hidden, communications 348
8.C.1 NTD with multiple channels (or high noise) 350
8.C.2 Stochastic Carrier Wave 352

8.D Problems 356

9 Machine-Learning based Computational Science 357
9.A Model Tuning Metaheuristics and Model Selection Methods 357
9.A.1 Gradient Ascent 358
9.A.2 Steepest Ascent Hill Climbing 359
9.A.3 Simulated Annealing 361
9.A.4 Taboo Search 362
9.A.5 Population-based metaheuristics 363
9.A.5.1 Population with evolution 363
9.A.5.2 Population with swarm intelligence 365
9.A.5.3 Indirect Interaction via Artifact 365
9.A.6 Problems 366

  9.B.1 Physics unifications and applications 367
  9.B.2 Physics & Statistics 369
  9.B.3 Inference via maximum entropy 371
  9.B.4 The distributions of nature via maximum entropy 372

9.C Conclusion: the fundamental role of machine-learning based optimization methods and representative in computational science and engineering 373

Bibliography 375

Index 399