CONTENTS

Introduction
BIOSAPIENS: A European Network of Excellence
to develop genome annotation resources (D. Frishman, A. Valencia) 1

SECTION 1: Gene definition

Chapter 1.1
State of the art in eukaryotic gene prediction (T. Alioto, R. Guigó) 7

1 Introduction 7
2 Classes of information 10
   2.1 Extrinsic information 10
   2.2 Intrinsic information 11
      2.2.1 Signals 11
      2.2.2 Content 14
   2.3 Conservation 15
3 Frameworks for integration of information 17
   3.1 Exon-chaining 17
   3.2 Generative models: Hidden Markov models 18
      3.2.1 Basic hidden Markov models 18
      3.2.2 Generalized hidden Markov models 21
      3.2.3 Generalized pair HMMs 21
      3.2.4 Phylo-HMMs or evolutionary HMMs 23
   3.3 Discriminative learning 24
      3.3.1 Support vector machines 24
      3.3.2 Semi-Markov conditional random fields 25
   3.4 Combiners 25
4 Training 26
5 Evaluation of gene prediction methods 27
   5.1 The basic tools 27
   5.2 Systematic evaluation 28
Chapter 1.2

1 Introduction 41
2 Quality control of gene predictions 42
   2.1 Principles of quality control 42
      2.1.1 Violation of some generally valid rules about proteins 42
         2.1.1.1 Conflict between the presence of extracellular Pfam-A domain(s) in a protein and the absence of appropriate sequence signals 43
         2.1.1.2 Conflict between the presence of extracellular and cytoplasmic Pfam-A domains in a protein and the absence of transmembrane segments 43
         2.1.1.3 Co-occurrence of nuclear and extracellular domains in a predicted multidomain protein 43
         2.1.1.4 Domain size deviation 44
      2.1.2 Violation of some generally valid rules of protein-coding genes 44
         2.1.2.1 Chimeric proteins parts of which are encoded by exons located on different chromosomes 44
3 Results 44
   3.1 Validation of the MisPred pipeline 44
   3.2 Errors detected by the MisPred tools in public databases 46
      3.2.1 Analysis of the TrEMBL section of UniProtKB 46
      3.2.2 Analysis of sequences predicted by the EnsEMBL and Gnomon gene prediction pipelines 47
4 Alternative interpretations of the results of MisPred analyses 50
   4.1 MisPred has a low false positive rate 50
   4.2 MisPred detects errors in gene prediction 50
4.3 MisPred detects “errors” of biological processes  51
4.4 MisPred discovers exceptions to generally valid rules  51

5 Conclusions  52

SECTION 2: Gene regulation and expression

Chapter 2.1
Evaluating the prediction of cis-acting regulatory elements in genome sequences (O. Sand, J. Valéry Turatsinze, J. van Helden)  55

1 Introduction  55
2 Transcription factor binding sites and motifs  58
3 Scanning a sequence with a position-specific scoring matrix  59
   3.1 Background probability  61
   3.2 Probability of a sequence segment given the motif  62
   3.3 Scanning profiles  63
4 Evaluating pattern matching results  64
   4.1 Evaluation statistics  64
   4.2 Accuracy profiles  66
   4.3 Avoiding circularity in the evaluation  67
   4.4 Why the statistics involving TN should be avoided  67
   4.5 Difficulties for the evaluation of pattern matching  68
5 Discovering motifs in promoter sequences  69
   5.1 Example of pattern discovery result  70
   5.2 Evaluation statistics  72
   5.3 Correctness of predicted motifs for a collection of annotated regulons  73
   5.4 Distributions of motif scores in positive and negative testing sets  77
   5.5 The Receiver Operating Characteristics (ROC) curve  81
   5.6 Using ROC curves to find optimal parameters  83
6 Methodological issues for evaluating pattern discovery  83
7 Good practices for evaluating predictive tools  84
   7.1 Use comprehensive data sets  85
   7.2 Think about your negative control  85
   7.3 Ensure neutrality  85
8 What has not been covered in this chapter  86
9 Materials  87
Chapter 2.2
A biophysical approach to large-scale protein-DNA binding data
(T. Manke, H. Roider, M. Vingron) 91

1 Binding site predictions 92
2 Affinity model {XE “affinity model, TRAP”} 95
3 Affinity statistics {XE “affinity statistics”} 99
4 Applications 101
5 Summary 102

Chapter 2.3
From gene expression profiling to gene regulation (R. Coulson, T. Manke, K. Palin, H. Roider, O. Sand, J. van Helden, E. Ukkonen, M. Vingron, A. Brazma) 105

1 Introduction 105
2 Generating sets of co-expressed genes 106
3 Finding putative regulatory regions using comparative genomics 109
4 Detecting common transcription factors for co-expressed gene sets 111
5 Combining transcription factor information 114
6 “De novo” prediction of transcription factor binding motifs 115

SECTION 3: Annotation and genetics

Chapter 3
Annotation, genetics and transcriptomics (R. Mott) 123

1 Introduction 123
2 Genetics and gene function 125
   2.1 Genetic association studies in humans 125
3 Use of animal models 128
4 Transcriptomics: gene expression microarrays 130
5 Gene annotation 132

SECTION 4: Functional annotation of proteins

Chapter 4.1
Resources for functional annotation (A. J. Bridge, A. Lise Veuthey, N. J. Mulder) 139

1 Introduction 139
2 Resources for functional annotation – protein sequence databases 140
Chapter 4.2

Annotating bacterial genomes (C. Médigue, A. Danchin) 165

1 Background 165
2 Global sequence properties 170
3 Identifying genomic objects 172
4 Functional annotation 174
5 A recursive view of genome annotation 176
Chapter 4.3
Data mining in genome annotation (I. Artamonova, S. Kramer, D. Frishman) 191

1 Introduction 191
2 An overview of large biological databases 193
   2.1 Manually curated vs. automatic databases 193
   2.2 Manually curated databases: the Swiss-Prot example 196
   2.3 Automatically generated databases: the PEDANT example 198
3 Data mining in genome annotation 200
   3.1 General remarks 200
   3.2 Supervised learning 201
   3.3 Unsupervised learning 201
   3.4 Clustering 202
   3.5 Association rule mining 203
4 Applying association rule mining to the Swiss-Prot database 205
5 Applying association rule mining to the PEDANT database 207
6 Conclusion 210

Chapter 4.4

1 Homologous and non-homologous sequence methods for assigning protein functions 213
   1.1 Introduction 213
   1.2 Homologs, orthologs, paralogs 216
   1.3 The HAMAP resource for the annotation of prokaryotic protein sequences and their orthologues 219
   1.4 CATH, Gene3D & GeMMA 222
   1.5 From SMART to STRING and STITCH: diverse tools for deducing function from sequence 228
1.6 General approaches for inheriting functions between homologous proteins 230
1.7 Non-homologous methods for predicting protein function from sequence 234

Chapter 4.5

1 Introduction to protein structure and function 239
2 FireDB and firestar – the prediction of functionally important residues 241
  2.1 Introduction 241
  2.2 FireDB 242
  2.3 Firestar 244
3 Modelling local function conservation in sequence and structure space for predicting molecular function 246
  3.1 Introduction 246
  3.2 Method 246
  3.3 Application 247
4 Structural templates for functional characterization 249
  4.1 Introduction 249
  4.2 Predicting protein function using structural templates 249
  4.3 FLORA method 250
5 An integrated pipeline for functional prediction 252
  5.1 Introduction 252
  5.2 The ProFunc server 253
    5.2.1 Sequence-based searches 254
    5.2.2 Structure-based searches 255
  5.3 Case studies 257
    5.3.1 Case study 1: published function identified 257
    5.3.2 Case study 2: function unclear 259
  5.4 Conclusion 259

Chapter 4.6
Harvesting the information from a family of proteins (B. Vroling, G. Vriend) 263

1 Introduction 263
  1.1 Information transfer 264
2 Molecular class-specific information systems 265
  2.1 G-protein-coupled receptors 266
Contents

3 Extracting information from sequences 267
   3.1 Correlated mutation analysis 268
4 Correlation studies on GPCRs 269
   4.1 Evolutionary trace method 271
   4.2 Entropy-variability analysis 273
   4.3 Sequence harmony 274
5 Discussion 274

SECTION 5: Protein structure prediction

Chapter 5.1
Structure prediction of globular proteins (A. Tramontano, D. Jones, L. Rychlewski, R. Casadio, P. Martelli, D. Raimondo and A. Giorgetti) 283

1 The folding problem 283
2 The evolution of protein structures and its implications for protein structure prediction 286
3 Template based modelling 287
   3.1 Homology-based selection of the template 288
   3.2 Fold recognition 288
   3.3 Using sequence based tools for selecting the template 289
   3.4 Completing and refining the model 291
   3.5 Current state of the art in template based methods 292
4 Template-free protein structure prediction 293
   4.1 Energy functions for protein structure prediction 296
   4.2 Lattice methods 297
   4.3 Fragment assembly methods 298
   4.4 Practical considerations 299
5 Automated structure prediction 300
   5.1 Practical lessons from benchmarking experiments 302
6 Conclusions and future outlook 304

Chapter 5.2
The state of the art of membrane protein structure prediction: from sequence to 3D structure (R. Casadio, P. Fariselli, P. L. Martelli, A. Pierleoni, I. Rossi, G. von Heijne) 309

1 Why membrane proteins? 309
2 Many functions 311
3 Bioinformatics and membrane proteins: is it feasible to predict the 3D structure of a membrane protein? 311

XIV
5.1 Constraints-based models of metabolism 343
   5.1.1 The flux balance hypothesis 343
   5.1.2 Modelling the growth medium 344
   5.1.3 Biomass function 345
5.2 Predicting metabolic capabilities 345
   5.2.1 Predicting growth on a defined medium 345
   5.2.2 Predicting gene essentiality 346
5.3 Assessing and correcting models using experimental data 347
5.4 Structural properties of the flux cone 347
5.5 Working with constraints-based models 348

Chapter 6.2

1 Introduction 353
2 Experimental methods 354
3 Protein interaction databases 356
4 Data standards for molecular interactions 356
5 The IntAct molecular interaction database 360
6 Interaction networks 362
7 Visualization software for molecular networks 365
8 Estimates of the number of protein interactions 371
9 Multi-protein complexes 372
10 Network modules 373
11 Diseases and protein interaction networks 376
12 Sequence-based prediction of protein interactions 380
   12.1 Phylogenetic profiling 381
   12.2 Similarity of phylogenetic trees 383
   12.3 Gene neighbourhood conservation 384
   12.4 Gene fusion 385
13 Integration of experimentally determined and predicted interactions 385
14 Domain–domain interactions 389
15 Biomolecular docking 395
   15.1 Protein-ligand docking 396
   15.2 Protein–protein docking 398

XVI
SECTION 7: Infrastructure for distributed protein annotation

Chapter 7
Infrastructure for distributed protein annotation (G. A. Reeves, A. Prlic, R. C. Jimenez, E. Kulesha, H. Hermjakob) 413

1 Introduction 413
2 The Distributed Annotation System (DAS) 415
3 DAS infrastructure 415
   3.1 DASTY2 – a protein sequence-oriented DAS client 418
   3.2 SPICE – a protein structure-oriented DAS client 418
   3.3 Ensembl 420
   3.4 DAS servers 422
4 The protein feature ontology 422
5 Conclusion 425

SECTION 8: Applications

Chapter 8.1
Viral bioinformatics (B. Adams, A. Carolyn McHardy, C. Lundegaard, T. Lengauer) 429

1 Introduction 429
2 Viral evolution in the human population 430
   2.1 Biology and genetics 430
   2.2 Vaccine strain selection for endemic influenza 431
   2.3 Pandemic influenza 433
   2.4 Conclusion 434
3 Interaction between the virus and the human immune system 434
   3.1 Introduction to the human immune system 434
   3.2 Epitopes 436
   3.3 Prediction of epitopes 437
   3.4 Epitope prediction in viral pathogens in a vaccine perspective 441
4 Viral evolution in the human host 442
   4.1 Introduction 442
   4.2 Replication cycle of HIV 443
   4.3 Targets for antiviral drug therapy 444
   4.4 Manual selection of antiretroviral combination drug therapies 444
   4.5 Data sets for learning viral resistance 445
   4.6 Computational procedures for predicting resistance 446
4.7 Clinical impact of bioinformatical resistance testing 449
4.8 Bioinformatical support for applying coreceptor inhibitors 450

5 Perspectives 450

Chapter 8.2

1 Introduction 453
2 Prediction of variant location 455
3 Prediction of variant function – analysis of the role of alternative splicing in changing function by modulation of functional residues 458
   3.1 Functions associated with alternative splicing 458
   3.2 Functional adaptation through alternative splicing 458
      3.2.1 Tafazzin 459
      3.2.2 Phosphoribosylglycinamide formyltransferase (GARS-AIRS-GART) 461
   3.3 Analysis across the ENCODE dataset 462
4 Prediction of variant structure 463
5 Summary of effects of alternative splicing 467
6 Prediction of principal isoforms 472
   6.1 A series of automatic methods for predicting the principal isoform 473
      6.1.1 Methods 474
      6.1.2 Evaluation of pipeline definitions 474
7 The ENCODE pipeline – an automated workflow for analysis of human splice isoforms 477
   7.1 Behind EPipe 477
   7.2 Example workflow: IFN alpha/beta receptor protein 479
   7.3 Future perspectives 480
Modern Genome Annotation
The Biosapiens Network
(Eds.) D. Frishman; A. Valencia
2008, 490 p. 135 illus., 110 in color., Hardcover
ISBN: 978-3-211-75122-0