

Golan Yona

Introduction to Computational Proteomics



Acknowledgments

I would like to thank the following people for their help in reviewing chapters of this book and providing useful comments: Ran El-Yaniv, Nir Kalisman, Klara Kedem, Danny Barash, Yoram Gdalyahu, Sergio Moreno, Peter Mirani, Dahlia Weiss, Adelene Sim, Assaf Oron and Fengzhu Sun.

Certain sections of this book are based on joint research with my former students, and I would like to thank them for their contributions: Itai Sharon, Liviu Popescu, Chin-Jen Ku, Niranjan Nagarajan, Helgi Ingolfsson, Umar Syed, Michael Quist, Richard Chung, Shafquat Rahman, William Dirks, Aaron Birkland, Paul Shafer and Timothy Isganitis.

To Michael Levitt, for his continuous encouragement and support over the years. To Jawahar Sudhamsu, for his kind help with many figures and his invaluable friendship.



Contents

I	The Basics	1
1	What Is Computational Proteomics?	3
1.1	The complexity of living organisms	3
1.2	Proteomics in the modern era	4
1.3	The main challenges in computational proteomics	5
1.3.1	Analysis of individual molecules	5
1.3.1.1	Sequence analysis	5
1.3.1.2	Structure analysis	6
1.3.2	From individual proteins to protein families	6
1.3.3	Protein classification, clustering and embedding	7
1.3.4	Interactions, pathways and gene networks	7
2	Basic Notions in Molecular Biology	9
2.1	The cell structure of organisms	9
2.2	It all starts from the DNA	10
2.3	Proteins	12
2.4	From DNA to proteins	15
2.5	Protein folding - from sequence to structure	18
2.6	Evolution and relational classes in the protein space	20
2.7	Problems	22
3	Sequence Comparison	23
3.1	Introduction	23
3.2	Alignment of sequences	24
3.2.1	Global sequence similarity	25
3.2.1.1	Calculating the global similarity score	26
3.2.2	Penalties for gaps	27
3.2.2.1	Linear gap functions	28
3.2.3	Local alignments	29
3.2.3.1	Calculating the local similarity score	30
3.3	Heuristic algorithms for sequence comparison	31
3.4	Probability and statistics of sequence alignments	32
3.4.1	Basic random model	33
3.4.2	Statistics of global alignment	33

3.4.2.1	Fixed alignment - global alignment without gaps	34
3.4.2.2	Optimal alignment	34
3.4.2.3	The zscore approach	35
3.4.3	Statistics of local alignments without gaps	37
3.4.3.1	Fixed alignment	37
3.4.3.2	Optimal alignment	38
3.4.4	Local alignments with gaps	43
3.4.5	Handling low-complexity sequences	45
3.4.6	Sequence identity and statistical significance	48
3.4.7	Similarity, homology and transitivity	49
3.5	Scoring matrices and gap penalties	50
3.5.1	Scoring matrices for nucleic acids	50
3.5.2	Scoring matrices for amino acids	50
3.5.2.1	The PAM family of scoring matrices	52
3.5.2.2	The BLOSUM family of scoring matrices	57
3.5.3	Information content of scoring matrices	58
3.5.3.1	Choosing the scoring matrix	60
3.5.4	Gap penalties	61
3.6	Distance and pseudo-distance functions for proteins	62
3.7	Further reading	66
3.8	Conclusions	68
3.9	Appendix - non-linear gap penalty functions	69
3.10	Appendix - implementation of BLAST and FASTA	72
3.10.1	FASTA	72
3.10.2	BLAST	72
3.11	Appendix - performance evaluation	75
3.11.1	Accuracy, sensitivity and selectivity	76
3.11.2	ROC	79
3.11.3	Setup and normalization	80
3.11.4	Reference datasets, negatives and positives	82
3.11.5	Training and testing algorithms	83
3.12	Appendix - basic concepts in probability	85
3.12.1	Probability mass and probability density	85
3.12.2	Moments	86
3.12.3	Conditional probability and Bayes' formula	87
3.12.4	Common probability distributions	88
3.12.5	The entropy function	91
3.12.6	Relative entropy and mutual information	92
3.12.7	Prior and posterior, ML and MAP estimators	93
3.12.8	Decision rules and hypothesis testing	95
3.13	Appendix - metrics and real normed spaces	98
3.14	Problems	100

4	Multiple Sequence Alignment, Profiles and Partial Order Graphs	105
4.1	Dynamic programming in N dimensions	106
4.1.1	Scoring functions	107
4.2	Classical heuristic methods	108
4.2.1	Star alignment	109
4.2.2	Tree alignment	110
4.3	MSA representation and scoring	113
4.3.1	The consensus sequence of an MSA	113
4.3.2	Regular expressions	114
4.3.3	Profiles and position-dependent scores	116
4.3.3.1	Generating a profile	116
4.3.3.2	Pseudo-counts	117
4.3.3.3	Weighting sequences	122
4.3.4	Position-specific scoring matrices	126
4.3.4.1	Using PSSMs with the dynamic programming algorithm	128
4.3.5	Profile-profile comparison	128
4.4	Iterative and progressive alignment	132
4.4.1	PSI-BLAST - iterative profile search algorithm	132
4.4.2	Progressive star alignment	136
4.4.3	Progressive profile alignment	137
4.5	Transitive alignment	138
4.5.1	T-coffee	139
4.6	Partial order alignment	141
4.6.1	The partial order MSA model	142
4.6.2	The partial order alignment algorithm	144
4.7	Further reading	148
4.8	Conclusions	149
4.9	Problems	150
5	Motif Discovery	155
5.1	Introduction	155
5.2	Model-based algorithms	156
5.2.1	The basic model	157
5.2.2	Model quality	158
5.2.2.1	Case 1: model unknown, patterns are given	159
5.2.2.2	Case 2: model is given, patterns are unknown	160
5.3	Searching for good models	160
5.3.1	The Gibbs sampling algorithm	161
5.3.1.1	Improvements	162
5.3.2	The MEME algorithm	162

	5.3.2.1	E-step	164
	5.3.2.2	M-step	165
	5.3.2.3	The iterative procedure	166
5.4		Combinatorial approaches	167
	5.4.1	Clique elimination	167
	5.4.2	Random projections	170
5.5		Further reading	173
5.6		Conclusions	175
5.7		Appendix - the Expectation-Maximization algorithm	176
5.8		Problems	180
6		Markov Models of Protein Families	183
6.1		Introduction	183
6.2		Markov models	184
	6.2.1	Gene prediction	184
	6.2.2	Formal definition	188
		6.2.2.1 Visible symbols and hidden Markov models	190
		6.2.2.2 The model's components	190
6.3		Main applications of hidden Markov models	191
	6.3.1	The evaluation problem	192
		6.3.1.1 The HMM forward algorithm	194
		6.3.1.2 The HMM backward algorithm	194
		6.3.1.3 Using HMMs for classification	196
	6.3.2	The decoding problem	196
	6.3.3	The learning problem	198
		6.3.3.1 The forward-backward algorithm	198
		6.3.3.2 Learning from multiple training sequences	200
	6.3.4	Handling machine precision limitations	201
	6.3.5	Constructing a model	202
		6.3.5.1 General model topology	202
		6.3.5.2 Model architecture	202
		6.3.5.3 Hidden Markov models for protein families	204
		6.3.5.4 Handling silent states	206
		6.3.5.5 Building a model from an MSA	206
		6.3.5.6 Single model vs. mixtures of multiple models	209
6.4		Higher order models, codes and compression	210
	6.4.1	Fixed order models	211
	6.4.2	Variable-order Markov models	213
		6.4.2.1 Codes and compression	214
		6.4.2.2 Compression and prediction	217
		6.4.2.3 Lempel-Ziv compression and extensions	218
		6.4.2.4 Probabilistic suffix trees	219

6.4.2.5	Sparse Markov transducers	227
6.4.2.6	Prediction by partial matches	229
6.5	Further reading	231
6.6	Conclusions	232
6.7	Problems	233
7	Classifiers and Kernels	235
7.1	Generative models vs. discriminative models	235
7.2	Classifiers and discriminant functions	237
7.2.1	Linear classifiers	238
7.2.2	Linearly separable case	241
7.2.3	Maximizing the margin	244
7.2.4	The non-separable case - soft margin	246
7.2.5	Non-linear discriminant functions	249
7.2.5.1	Mercer kernels	253
7.3	Applying SVMs to protein classification	255
7.3.1	String kernels	256
7.3.1.1	Simple string kernel - the spectrum kernel	256
7.3.1.2	The mismatch spectrum kernel	257
7.3.2	The pairwise kernel	257
7.3.3	The Fischer kernel	258
7.3.4	Mutual information kernels	259
7.4	Decision trees	262
7.4.1	The basic decision tree model	263
7.4.2	Training decision trees	264
7.4.2.1	Impurity measures for multi-valued attributes	267
7.4.2.2	Missing attributes	268
7.4.2.3	Tree pruning	268
7.4.3	Stochastic trees and mixture models	270
7.4.4	Evaluation of decision trees	272
7.4.4.1	Handling skewed distributions	274
7.4.5	Representation and feature extraction	275
7.4.5.1	Feature processing	276
7.4.5.2	Dynamic attribute filtering	277
7.4.5.3	Binary splitting	278
7.5	Further reading	279
7.6	Conclusions	280
7.7	Appendix - estimating the significance of a split	281
7.8	Problems	288

8	Protein Structure Analysis	291
8.1	Introduction	291
8.2	Structure prediction - the protein folding problem	293
8.2.1	Protein secondary structure prediction	296
8.2.1.1	Secondary structure assignment	297
8.2.1.2	Secondary structure prediction	299
8.2.1.3	Accuracy of secondary structure prediction	301
8.3	Structure comparison	303
8.3.1	Algorithms based on inter-atomic distances	305
8.3.1.1	The RMSd measure	305
8.3.1.2	The structural algorithm	309
8.3.1.3	The URMS distance	311
8.3.1.4	The URMS-RMS algorithm	312
8.3.2	Distance matrix based algorithms	318
8.3.2.1	Dali	319
8.3.2.2	CE	322
8.3.3	Geometric hashing	324
8.3.4	Statistical significance of structural matches	327
8.3.5	Evaluation of structure comparison	330
8.4	Generalized sequence profiles - integrating secondary structure with sequence information	332
8.5	Further reading	336
8.6	Conclusions	339
8.7	Appendix - minimizing RMSd	340
8.8	Problems	342
9	Protein Domains	345
9.1	Introduction	345
9.2	Domain detection	348
9.2.1	Domain prediction from 3D structure	349
9.2.2	Domain analysis based on predicted measures of structural stability	351
9.2.3	Domain prediction based on sequence similarity search	355
9.2.4	Domain prediction based on multiple sequence alignments	361
9.3	Learning domain boundaries from multiple features	364
9.3.1	Feature optimization	365
9.3.2	Scaling features	366
9.3.3	Post-processing predictions	366
9.3.4	Training and evaluation of models	369
9.4	Testing domain predictions	370
9.4.1	Selecting more likely partitions	373
9.4.1.1	Computing the prior $P(D)$	375

9.4.1.2	Computing the likelihood $P(S D)$	376
9.4.2	The distribution of domain lengths	378
9.5	Multi-domain architectures	380
9.5.1	Hierarchies of multi-domain proteins	380
9.5.2	Relationships between domain architectures	381
9.5.3	Semantically significant domain architectures	385
9.6	Further reading	387
9.7	Conclusions	389
9.8	Appendix - domain databases	390
9.9	Problems	393

II Putting All the Pieces Together 395

10	Clustering and Classification	397
10.1	Introduction	397
10.2	Clustering methods	399
10.3	Vector-space clustering algorithms	401
10.3.1	The k-means algorithm	402
10.3.2	Fuzzy clustering	404
10.3.3	Hierarchical algorithms	408
10.3.3.1	Hierarchical k-means	409
10.3.3.2	The statistical mechanics approach	409
10.4	Graph-based clustering algorithms	410
10.4.1	Pairwise clustering algorithms	411
10.4.1.1	The single linkage algorithm	412
10.4.1.2	The complete linkage algorithm	414
10.4.1.3	The average linkage algorithm	414
10.4.2	Collaborative clustering	415
10.4.3	Spectral clustering algorithms	421
10.4.4	Markovian clustering algorithms	425
10.4.5	Super-paramagnetic clustering	427
10.5	Cluster validation and assessment	428
10.5.1	External indices of validity	430
10.5.1.1	The case of known classification	430
10.5.1.2	The case of known relations	433
10.5.2	Internal indices of validity	434
10.5.2.1	The MDL principle	434
10.5.2.2	Cross-validation	439
10.6	Clustering proteins	440
10.6.1	Domains vs. complete proteins	440
10.6.2	Graph representation	441
10.6.3	Graph-based protein clustering	442
10.6.4	Integrating multiple similarity measures	444

10.7	Further reading	448
10.8	Conclusions	450
10.9	Appendix - cross-validation tests	451
10.10	Problems	457
11	Embedding Algorithms and Vectorial Representations	459
11.1	Introduction	459
11.2	Structure preserving embedding	461
11.2.1	Maximal variance embeddings	461
11.2.1.1	Principal component analysis	462
11.2.1.2	Singular value decomposition	467
11.2.2	Distance preserving embeddings	467
11.2.2.1	Multidimensional scaling	468
11.2.2.2	Embedding through random projections	474
11.2.3	Manifold learning - topological embeddings	478
11.2.3.1	Embedding with geodesic distances	479
11.2.3.2	Preserving local neighborhoods	482
11.2.3.3	Distributional scaling	484
11.3	Setting the dimension of the host space	488
11.4	Vectorial representations	490
11.4.1	Internal representations	492
11.4.2	Collective and external representations	493
11.4.2.1	Choosing a reference set and an association measure	494
11.4.2.2	Transformations and normalizations	495
11.4.2.3	Noise reduction	495
11.4.2.4	Comparing distance profiles	496
11.4.2.5	Distance profiles and mixture models	500
11.5	Further reading	502
11.6	Conclusions	503
11.7	Problems	504
12	Analysis of Gene Expression Data	505
12.1	Introduction	505
12.2	Microarrays	509
12.2.1	Datasets	512
12.3	Analysis of individual genes	513
12.4	Pairwise analysis	515
12.4.1	Measures of expression similarity	517
12.4.1.1	Shifts	520
12.4.2	Missing data	521
12.4.3	Correlation vs. anti-correlation	523
12.4.4	Statistical significance of expression similarity	524

12.4.5	Evaluating similarity measures	527
12.4.5.1	Estimating baseline performance	528
12.5	Cluster analysis and class discovery	529
12.5.1	Validating clustering results	534
12.5.2	Assessing individual clusters	536
12.5.3	Enrichment analysis	538
12.5.3.1	The gene ontology	538
12.5.3.2	Gene set enrichment	541
12.5.4	Limitations of mRNA arrays	544
12.6	Protein arrays	545
12.6.1	Mass-spectra data	546
12.7	Further reading	548
12.8	Conclusions	550
12.9	Problems	551
13	Protein-Protein Interactions	553
13.1	Introduction	553
13.2	Experimental detection of protein interactions	556
13.2.1	Traditional methods	557
13.2.1.1	Affinity chromatography	557
13.2.1.2	Co-immunoprecipitation	558
13.2.2	High-throughput methods	558
13.2.2.1	The two-hybrid system	558
13.2.2.2	Tandem affinity purification	560
13.2.2.3	Protein arrays	561
13.3	Prediction of protein-protein interactions	561
13.3.1	Structure-based prediction of interactions	562
13.3.1.1	Protein docking and prediction of interaction sites	563
13.3.1.2	Extensions to sequences of unknown structures	567
13.3.2	Sequence-based inference	568
13.3.2.1	Gene preservation and locality	568
13.3.2.2	Co-evolution analysis	571
13.3.2.3	Predicting the interaction interface	578
13.3.2.4	Sequence signatures and domain-based prediction	582
13.3.3	Gene co-expression	589
13.3.4	Hybrid methods	589
13.3.5	Training and testing models on interaction data	591
13.4	Interaction networks	592
13.4.1	Topological properties of interaction networks	593
13.4.2	Applications	601

13.4.3	Network motifs and the modular organization of networks	602
13.5	Further reading	605
13.6	Conclusions	607
13.7	Appendix - DNA amplification and protein expression . . .	608
13.7.1	Plasmids	608
13.7.2	SDS-PAGE	608
13.8	Appendix - the Pearson correlation	610
13.8.1	Uneven divergence rates	610
13.8.2	Insensitivity to the size of the dataset	610
13.8.3	The effect of outliers	611
13.9	Problems	613
14	Cellular Pathways	615
14.1	Introduction	615
14.2	Metabolic pathways	618
14.3	Pathway prediction	621
14.3.1	Metabolic pathway prediction	621
14.3.2	Pathway prediction from blueprints	623
14.3.2.1	The problem of pathway holes	623
14.3.2.2	The problem of ambiguity	623
14.3.3	Expression data and pathway analysis	624
14.3.3.1	Deterministic gene assignments	626
14.3.3.2	Fuzzy assignments	629
14.3.4	From model to practice	632
14.4	Regulatory networks: modules and regulation programs . .	635
14.5	Pathway networks and the minimal cell	640
14.6	Further reading	642
14.7	Conclusions	645
14.8	Problems	646
15	Learning Gene Networks with Bayesian Networks	649
15.1	Introduction	649
15.1.1	The basics of Bayesian networks	650
15.2	Computing the likelihood of observations	654
15.3	Probabilistic inference	655
15.3.1	Inferring the values of variables in a network	656
15.3.2	Inference of multiple unknown variables	660
15.4	Learning the parameters of a Bayesian network	661
15.4.1	Computing the probability of new instances	666
15.4.2	Learning from incomplete data	667
15.5	Learning the structure of a Bayesian network	669
15.5.1	Alternative score functions	672

15.5.2	Searching for optimal structures	674
15.5.2.1	Greedy search	675
15.5.2.2	Sampling techniques	675
15.5.2.3	Model averaging	676
15.5.3	Computing the probability of new instances	678
15.6	Learning Bayesian networks from microarray data	678
15.7	Further reading	682
15.8	Conclusions	683
15.9	Problems	684
	References	687
	Conference Abbreviations	733
	Acronyms	735
	Index	738



Preface

Computational molecular biology, or simply **computational biology**, is a term generally used to describe a broad set of techniques, models and algorithms that are applied to problems in biology. This is a relatively new discipline that is rooted in two different disciplines: computer science and molecular biology. Being on the border line between the two disciplines, it is related to fields of intensive research in both. The goal of this book is to introduce the field of computational biology through a focused approach that tackles the different steps and problems involved with protein analysis, classification and meta-organization. Of special interest are problems related to the study of protein-based cellular networks. All these tasks constitute what is referred to as **computational proteomics**.

This is a broad goal, and indeed the book covers a variety of topics. The first part covers methods to identify the building blocks of the protein space, such as motifs and domains, and algorithms to assess similarity between proteins. This includes sequence and structure analysis, and mathematical models (such as hidden Markov models and support vector machines) that are used to represent protein families and classify new instances. The second part covers methods that explore higher order structure in the protein space, through the application of unsupervised learning algorithms, such as clustering and embedding. The third part discusses methods that explore and unravel the broader context of proteins, such as prediction of interactions with other molecules, transcriptional regulation and reconstruction of cellular pathways and gene networks.

The book is structured also based on the type of the biological data analyzed. It starts with the analysis of individual entities, and works its way up through the analysis of more complex entities. The first chapters provide a brief introduction to the molecular biology of the main entities that are of interest when studying the protein space, and an overview of the main problems we will focus on. These are followed by a chapter on pairwise sequence alignment, including rigorous and heuristic algorithms, and statistical assessment of sequence similarity. Next we discuss algorithms for multiple sequence alignment, as well as generative and discriminative models of protein families. We proceed to discuss motif detection, domain prediction and protein structure analysis. All these algorithms and models are elemental to the methods that are discussed in the next couple of chapters on clustering, embedding and protein classification. The last several chapters are devoted to the analysis of the broader biological context of proteins, which is essential to fully and ac-

curately characterize proteins and their cellular counterparts. This includes gene expression analysis, prediction and analysis of protein-protein interactions, and the application of probabilistic models to study pathways, gene networks and causality in cells.

The book is intended for computer scientists, statisticians, mathematicians and biologists. The goal of this book is to provide a coherent view of the field and the main problems involved with the analysis of complex biological systems and specifically the protein space. It offers rigorous and formal descriptions, when possible, with detailed algorithmic solutions and models. Each chapter is followed by problem sets from courses the author has taught at Cornell University and at the Technion, with emphasis on a practical approach. Basic background in probability and statistics is assumed, but is also provided in an appendix to Chapter 3. Knowledge of molecular biology is not required, but we highly recommend referring to a specialized book in molecular biology or biochemistry for further information (for a list of recommended books, see the book's website at biozon.org/proteomics/)

It should be noted that the interaction of computer science and molecular biology as embodied in computational biology is not a one way street. In this book we focus on algorithms and models and their application to biological problems. The opposite scenario, where biological systems are used to solve mathematical problems (as in DNA computing), is also of interest; however it is outside the scope of this book. Nevertheless, it is fascinating to see how biology affects the way we think, by introducing new concepts and new models of computation (well known examples include neural networks and genetic algorithms). This interaction invigorates fields like statistics and computer science and triggers the development of new models and algorithms that have a great impact on other fields of science as well.

Before we start, we should mention the term **Bioinformatics**, which is equivalent to computational biology. Some make a distinction and use the term computational biology to refer to the *development* of novel algorithms and models to solve biological problems, while Bioinformatics is used to refer to the *application* of these algorithms to biological data. However, this difference in semantics is somewhat fuzzy, and practically the terms are used interchangeably.