

METHODS EXPRESS

# Bioinformatics

edited by Paul H. Dear



# Contents

|                  |      |
|------------------|------|
| Contributors     | x    |
| Preface          | xii  |
| Acknowledgements | xii  |
| Before you begin | xiii |
| Abbreviations    | xv   |

|                      |             |
|----------------------|-------------|
| <b>Color section</b> | <b>xvii</b> |
|----------------------|-------------|

## **Chapter 1. Database resources for wet-bench scientists**

*Neil Hall and Lynn M. Schriml*

|                                 |    |
|---------------------------------|----|
| 1. Introduction                 | 1  |
| 1.1 Types of databases          | 1  |
| 1.2 Database resources at NCBI  | 2  |
| 2. Methods and approaches       | 4  |
| 2.1 Searching databases at NCBI | 4  |
| 2.2 Downloading NCBI datasets   | 11 |
| 3. Troubleshooting              | 11 |
| 4. Additional web resources     | 12 |
| 5. References                   | 13 |

## **Chapter 2. Navigating sequenced genomes**

*Melody S. Clark and Thomas Schlitt*

|                                                            |    |
|------------------------------------------------------------|----|
| 1. Introduction                                            | 15 |
| 2. Methods and approaches                                  | 16 |
| 2.1 Finding genome resources for an organism               | 16 |
| 2.2 Browsing vertebrate genomes with Ensembl               | 18 |
| 2.3 Integr8 – an Ensembl lookalike for microbes            | 22 |
| 2.4 Other web-based genome browsers                        | 24 |
| 2.5 Specialized sites                                      | 26 |
| 2.6 Downloading data with BioMart                          | 27 |
| 2.7 Browsing genomes 'off line' using stand-alone software | 30 |
| 2.8 Linking your own data to a genome browser              | 33 |
| 3. References                                              | 38 |

**Chapter 3. Sequence similarity searches***Jaap Heringa and Walter Pirovano*

|                                                                  |    |
|------------------------------------------------------------------|----|
| 1. Introduction                                                  | 39 |
| 1.1 Comparative sequence analysis                                | 39 |
| 1.2 Sequence alignment as a reflection of similarity             | 39 |
| 1.3 Similarity versus homology                                   | 40 |
| 1.4 Techniques for pairwise alignment                            | 41 |
| 1.5 Alignment scores as a measure of similarity                  | 42 |
| 1.6 Sequence identity as a measure of similarity                 | 43 |
| 1.7 Statistics of alignment similarity scores                    | 43 |
| 1.8 Protein domains                                              | 44 |
| 2. Methods and approaches                                        | 44 |
| 2.1 Should one compare protein or nucleotide sequences?          | 45 |
| 2.2 Curated and annotated sequence databases                     | 46 |
| 2.3 Heuristic sequence similarity searching methods              | 47 |
| 2.4 Statistical significance of search results – <i>E</i> values | 56 |
| 2.5 Fast Smith–Waterman local alignment searches                 | 59 |
| 2.6 Profile searching                                            | 60 |
| 3. Troubleshooting                                               | 65 |
| 3.1 Iterative homology searching problems                        | 65 |
| 3.2 Post-processing of homology searches                         | 66 |
| 3.3 Evaluating sequence database searches                        | 66 |
| 4. References                                                    | 67 |

**Chapter 4. Gene prediction***Marie-Adele Rajandream*

|                                  |     |
|----------------------------------|-----|
| 1. Introduction                  | 71  |
| 1.1 <i>Ab initio</i> methods     | 72  |
| 1.2 Comparative methods          | 73  |
| 2. Methods and approaches        | 74  |
| 2.1 Predicting eukaryotic genes  | 75  |
| 2.2 Predicting prokaryotic genes | 90  |
| 3. Troubleshooting               | 98  |
| 4. Additional web resources      | 99  |
| 5. References                    | 101 |

**Chapter 5. Prediction of noncoding transcripts***Alex Bateman and Sam Griffiths-Jones*

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| 1. Introduction                                                   | 103 |
| 2. Methods and approaches                                         | 105 |
| 2.1 <i>Ab initio</i> versus family-specific searches              | 105 |
| 2.2 Web servers for the detection of single, specific RNA classes | 106 |
| 2.3 Web servers for the prediction of multiple RNA classes        | 111 |

|                                         |     |
|-----------------------------------------|-----|
| 3. Troubleshooting                      | 114 |
| 3.1 RNA-derived repeats and pseudogenes | 114 |
| 3.2 Computational complexity            | 115 |
| 4. References                           | 115 |

## **Chapter 6. Finding regulatory elements in DNA sequence**

*Debraj GuhaThakurta and Gary D. Stormo*

|                                                                                                  |     |
|--------------------------------------------------------------------------------------------------|-----|
| 1. Introduction                                                                                  | 117 |
| 1.1 Background                                                                                   | 117 |
| 1.2 An overview of progress in the computational identification of DNA sequence motifs           | 118 |
| 1.3 Modeling and representation of DNA motifs                                                    | 119 |
| 2. Methods and approaches                                                                        | 123 |
| 2.1 Searching DNA for known motifs                                                               | 123 |
| 2.2 Discovery of DNA motifs from input DNA sequences                                             | 126 |
| 2.3 Comparative genomics and phylogenetic footprinting in the search for DNA regulatory elements | 132 |
| 2.4 Composite DNA motifs and <i>cis</i> -regulatory modules                                      | 134 |
| 3. Additional web resources                                                                      | 135 |
| 4. References                                                                                    | 136 |

## **Chapter 7. Expressed sequence tags**

*Arthur Gruber*

|                                                         |     |
|---------------------------------------------------------|-----|
| 1. Introduction                                         | 141 |
| 1.1 EST library construction and sequencing             | 142 |
| 1.2 Representation: normalized and subtracted libraries | 144 |
| 2. Methods and approaches                               | 145 |
| 2.1 Overview                                            | 145 |
| 2.2 EST databases                                       | 146 |
| 2.3 Automated EST pre-processing pipelines              | 150 |
| 2.4 Transcript reconstruction                           | 155 |
| 2.5 Redundancy estimation                               | 160 |
| 2.6 Electronic gene expression profiles                 | 162 |
| 2.7 Mapping ESTs to the genome                          | 162 |
| 3. Troubleshooting                                      | 163 |
| 3.1 Clone chimerism                                     | 163 |
| 3.2 SNPs                                                | 164 |
| 3.3 Repeat masking                                      | 164 |
| 3.4 Contamination                                       | 164 |
| 4. Additional web resources                             | 164 |
| 5. References                                           | 165 |

## **Chapter 8. Protein structure, classification, and prediction**

*Arthur M. Lesk*

|                                        |     |
|----------------------------------------|-----|
| 1. Introduction                        | 169 |
| 1.1 The chemical structure of proteins | 170 |

|     |                                                       |     |
|-----|-------------------------------------------------------|-----|
| 1.2 | The hierarchical form of protein architecture         | 172 |
| 1.3 | Domains                                               | 173 |
| 2.  | Methods and approaches                                | 173 |
| 2.1 | Accessing macromolecular structures on the web        | 173 |
| 2.2 | Classification of protein structures                  | 176 |
| 2.3 | Structural genomics                                   | 180 |
| 2.4 | Approaches to protein structure prediction            | 180 |
| 2.5 | Specialized methods for particular types of structure | 186 |
| 3.  | References                                            | 194 |

## Chapter 9. Gene ontology

*Vineet Sangar*

|     |                                                     |     |
|-----|-----------------------------------------------------|-----|
| 1.  | Introduction                                        | 195 |
| 1.1 | Gene ontology                                       | 196 |
| 1.2 | Structure of the GO database                        | 196 |
| 1.3 | The three GO ontologies                             | 198 |
| 1.4 | GO terms                                            | 199 |
| 1.5 | Evidence codes                                      | 199 |
| 2.  | Methods and approaches                              | 200 |
| 2.1 | GO browsers                                         | 200 |
| 2.2 | GO annotation tools                                 | 204 |
| 2.3 | Gene expression tools                               | 205 |
| 2.4 | Integration of GO with other classification systems | 206 |
| 3.  | Additional web resources                            | 206 |
| 4.  | References                                          | 207 |

## Chapter 10. Prediction of protein function

*Rodrigo Lopez*

|      |                                                                           |     |
|------|---------------------------------------------------------------------------|-----|
| 1.   | Introduction                                                              | 209 |
| 2.   | Methods and approaches                                                    | 210 |
| 2.1  | Required tools                                                            | 210 |
| 2.2  | Prediction and determination of physicochemical properties of proteins    | 210 |
| 2.3  | Determination of secondary structure from sequence                        | 215 |
| 2.4  | Determination of functional domains using pattern-matching methods        | 224 |
| 2.5  | Advanced methods combining several protein function prediction algorithms | 230 |
| 2.6  | Protein function prediction by transfer of annotation                     | 233 |
| 2.7  | Multiple sequence alignments and secondary databases                      | 234 |
| 2.8  | An overview of InterPro and CDD                                           | 235 |
| 2.9  | Recent advances in protein function prediction                            | 238 |
| 2.10 | Concluding remarks                                                        | 241 |
| 3.   | Additional web resources                                                  | 241 |
| 4.   | References                                                                | 242 |

**Chapter 11. Multiple sequence alignment***Burkhard Morgenstern*

|                                                                |     |
|----------------------------------------------------------------|-----|
| 1. Introduction                                                | 245 |
| 2. Methods and approaches                                      | 246 |
| 2.1 The alignment problem in computational biology             | 246 |
| 2.2 Pairwise sequence alignment                                | 247 |
| 2.3 Multiple sequence alignment                                | 249 |
| 2.4 Benchmarking and evaluation of multiple-alignment software | 250 |
| 2.5 Visualization and comparison of multiple alignments        | 251 |
| 2.6 Multiple alignment of large genomic sequences              | 251 |
| 2.7 Software tools for multiple alignment                      | 252 |
| 3. Additional web resources                                    | 262 |
| 4. References                                                  | 263 |

**Chapter 12. Inferring phylogenetic relationships from sequence data***Peter G. Foster*

|                                          |     |
|------------------------------------------|-----|
| 1. Introduction                          | 265 |
| 2. Methods and approaches                | 269 |
| 2.1 Alignments                           | 269 |
| 2.2 File formats                         | 269 |
| 2.3 Software                             | 270 |
| 2.4 Tree-building methods                | 271 |
| 2.5 Choosing a model                     | 274 |
| 2.6 A Bayesian approach to phylogenetics | 278 |
| 3. Troubleshooting                       | 280 |
| 4. References                            | 281 |

**Appendix**

|                                            |     |
|--------------------------------------------|-----|
| Additional useful bioinformatics resources | 283 |
|--------------------------------------------|-----|

|       |     |
|-------|-----|
| Index | 287 |
|-------|-----|