

---

# CONTENTS

|                                            |           |
|--------------------------------------------|-----------|
| LIST OF FIGURES                            | xi        |
| LIST OF TABLES                             | xv        |
| PREFACE                                    | xvii      |
| <b>1 INTRODUCTION</b>                      | <b>1</b>  |
| 1.1 The Human Genome Project               | 1         |
| 1.1.1 Origin                               | 2         |
| 1.1.2 Goals                                | 3         |
| 1.1.3 Benefits                             | 4         |
| 1.2 Informatics                            | 6         |
| 1.2.1 Data management                      | 6         |
| 1.2.2 Data analysis                        | 10        |
| 1.3 Summary                                | 12        |
| <b>2 BIOLOGICAL BACKGROUND</b>             | <b>15</b> |
| 2.1 Introduction                           | 15        |
| 2.2 Types of biological sequences          | 17        |
| 2.3 DNA extraction and amplification       | 21        |
| 2.3.1 Preparation                          | 21        |
| 2.3.2 Amplification                        | 25        |
| 2.4 Sequencing technology                  | 32        |
| 2.4.1 Chemical degradation                 | 33        |
| 2.4.2 Dideoxy chain termination            | 35        |
| 2.4.3 Large fragment sequencing strategies | 39        |
| 2.4.4 Large scale sequencing automation    | 40        |

|          |                                                                |     |
|----------|----------------------------------------------------------------|-----|
| 2.5      | Sequence databases                                             | 41  |
| 2.6      | Summary                                                        | 48  |
| <b>3</b> | <b>SEQUENCE ANALYSIS ALGORITHMS</b>                            | 51  |
| 3.1      | Introduction                                                   | 51  |
| 3.2      | Substitution score matrices                                    | 52  |
| 3.2.1    | Amino acid substitution matrix                                 | 52  |
| 3.2.2    | Nucleic acid substitution matrix                               | 54  |
| 3.3      | Pairwise sequence alignment                                    | 54  |
| 3.3.1    | Calculating alignment score                                    | 55  |
| 3.3.2    | Constructing alignment                                         | 62  |
| 3.4      | Multiple sequence alignment                                    | 67  |
| 3.4.1    | Simultaneous alignment                                         | 68  |
| 3.4.2    | Clustering strategies                                          | 70  |
| 3.4.3    | Randomized strategies                                          | 78  |
| 3.5      | Sequence similarity searching                                  | 85  |
| 3.5.1    | FASTA                                                          | 86  |
| 3.5.2    | BLAST                                                          | 89  |
| 3.5.3    | FLASH                                                          | 91  |
| 3.6      | Summary                                                        | 96  |
| <b>4</b> | <b>HIGH PERFORMANCE COMPUTING ARCHITECTURES AND TECHNIQUES</b> | 99  |
| 4.1      | Introduction                                                   | 99  |
| 4.2      | Parallelism: what and why?                                     | 100 |
| 4.3      | Types of computation                                           | 101 |
| 4.4      | Classification schemes                                         | 103 |
| 4.5      | Metrics                                                        | 106 |
| 4.6      | Performance related factors                                    | 107 |
| 4.7      | Summary                                                        | 109 |
| <b>5</b> | <b>MULTIPROCESSOR SEQUENCE ALIGNMENT</b>                       | 111 |
| 5.1      | Introduction                                                   | 111 |
| 5.2      | Pairwise sequence alignment                                    | 112 |
| 5.2.1    | Calculating the alignment score                                | 112 |

|          |                                                      |     |
|----------|------------------------------------------------------|-----|
| 5.2.2    | Constructing the sequence alignment                  | 119 |
| 5.3      | Multiple sequence alignment                          | 127 |
| 5.3.1    | Clustering strategies                                | 127 |
| 5.3.2    | Randomized strategies                                | 128 |
| 5.4      | Summary                                              | 141 |
| <b>6</b> | <b>MULTIPROCESSOR SEQUENCE SIMILARITY SEARCHING</b>  | 143 |
| 6.1      | Introduction                                         | 143 |
| 6.2      | Fine-grain approach                                  | 144 |
| 6.3      | Coarse-grain approach                                | 148 |
| 6.4      | Summary                                              | 156 |
| <b>7</b> | <b>BIOLOGICAL SEQUENCE RESOURCES ON THE INTERNET</b> | 159 |
| 7.1      | Introduction                                         | 159 |
| 7.2      | Accessing tools                                      | 160 |
| 7.2.1    | Email                                                | 160 |
| 7.2.2    | Usenet                                               | 163 |
| 7.2.3    | ListServers                                          | 166 |
| 7.2.4    | FTP                                                  | 168 |
| 7.2.5    | Telnet                                               | 170 |
| 7.2.6    | Gopher                                               | 172 |
| 7.2.7    | World Wide Web                                       | 176 |
| 7.3      | Search engines                                       | 181 |
| 7.3.1    | Netfind and ph                                       | 181 |
| 7.3.2    | Archie                                               | 183 |
| 7.3.3    | Veronica                                             | 185 |
| 7.3.4    | WWW search engines                                   | 186 |
| 7.4      | Summary                                              | 188 |
| <b>8</b> | <b>LOOKING TO THE FUTURE</b>                         | 191 |
|          | <b>REFERENCES</b>                                    | 195 |
|          | <b>INDEX</b>                                         | 209 |