

# Table des matières

|   |    |
|---|----|
| <b>Avant-propos</b> . . . . .   | 1  |
| Annie CHATEAU et Mikaël SALSON  |    |
| <br>  |    |
| <b>Présentation des auteurs</b> . . . . .   | 7  |
| <br>  |    |
| <b>Chapitre 1. Concepts méthodologiques : résolution algorithmique des problèmes bioinformatiques</b> . . . . . | 11 |
| Annie CHATEAU et Tom DAVOT-GRANGÉ   |    |
| 1.1. Données, modèles, formalisme des problèmes en bioinformatique . . . . .                                    | 11 |
| 1.1.1. Données . . . . .  | 11 |
| 1.1.2. Modélisation des génomes . . . . .   | 14 |
| 1.1.3. Problèmes en bioinformatique . . . . .   | 15 |
| 1.2. Préliminaires mathématiques . . . . .  | 15 |
| 1.2.1. Préliminaires sur la logique propositionnelle . . . . .  | 15 |
| 1.2.2. Préliminaires sur les ensembles . . . . .  | 16 |
| 1.3. Vocabulaire de l’algorithmique du texte . . . . .  | 18 |
| 1.4. Théorie des graphes . . . . .  | 20 |
| 1.4.1. Sous-graphes . . . . .   | 20 |
| 1.4.2. Cheminement dans un graphe . . . . .   | 22 |
| 1.4.3. Couplage . . . . .   | 23 |
| 1.4.4. Planarité . . . . .  | 23 |
| 1.4.5. Décomposition arborescente . . . . .   | 24 |
| 1.5. Problèmes algorithmiques . . . . .   | 25 |
| 1.5.1. Définition . . . . .   | 25 |
| 1.5.2. Problème de graphes . . . . .  | 26 |
| 1.5.3. Problème de satisfaisabilité . . . . .   | 28 |
| 1.6. Résolution des problèmes . . . . .   | 29 |

|   |    |
|---|----|
| 1.6.1. Algorithme . . . . .                       | 29 |
| 1.6.2. Complexité . . . . .                       | 30 |
| 1.6.3. Temps d'exécution . . . . .                | 33 |
| 1.7. Classes de complexité . . . . .              | 34 |
| 1.7.1. Généralité . . . . .                       | 35 |
| 1.7.2. Algorithmes exacts . . . . .               | 37 |
| 1.7.3. Algorithmes approchés . . . . .            | 40 |
| 1.7.4. Solveurs . . . . .                         | 42 |
| 1.8. Quelques techniques algorithmiques . . . . . | 43 |
| 1.8.1. Programmation dynamique . . . . .          | 43 |
| 1.8.2. Parcours d'arbres . . . . .                | 45 |
| 1.9. Validation . . . . .                         | 48 |
| 1.9.1. Différents types d'erreurs . . . . .       | 49 |
| 1.9.2. Mesures de qualité . . . . .               | 51 |
| 1.9.3. Et dans le cas non binaire ? . . . . .     | 54 |
| 1.10. Conclusion . . . . .                        | 54 |
| 1.11. Bibliographie . . . . .                     | 54 |

## **Chapitre 2. Indexation de séquences . . . . . 57**

Thierry LECROQ et Mikaël SALSON

|  |    |
|--|----|
| 2.1. Introduction . . . . .  | 57 |
| 2.1.1. L'indexation . . . . .  | 58 |
| 2.1.2. Quand indexer ? . . . . .   | 58 |
| 2.1.3. Qu'indexer ? . . . . .  | 59 |
| 2.1.4. Structures d'indexation et requêtes considérées . . . . .                       | 60 |
| 2.1.5. Notions de base et vocabulaire . . . . .  | 61 |
| 2.2. Indexation de mots . . . . .  | 62 |
| 2.2.1. Filtre de Bloom . . . . .   | 62 |
| 2.2.2. Liste inversée . . . . .  | 64 |
| 2.2.3. Graphes de De Bruijn . . . . .  | 67 |
| 2.2.4. Des structures efficaces pour des requêtes ciblées . . . . .                    | 69 |
| 2.3. Indexation plein texte . . . . .  | 69 |
| 2.3.1. Arbre des suffixes . . . . .  | 69 |
| 2.3.2. Table (étendue) des suffixes . . . . .  | 71 |
| 2.3.3. Transformée de Burrows-Wheeler . . . . .  | 74 |
| 2.4. Critères de choix d'indexation . . . . .  | 83 |
| 2.4.1. En fonction du type de requête nécessaire . . . . .                             | 83 |
| 2.4.2. En fonction du compromis espace-temps et de la quantité<br>de données . . . . . | 84 |

|  |            |
|--|------------|
| 2.4.3. En fonction de la nécessité d'ajouter ou de modifier<br>les données indexées . . . . .  | 85         |
| 2.4.4. Des choix d'indexation selon les applications . . . . .                                 | 86         |
| 2.5. Conclusions et perspectives . . . . .   | 88         |
| 2.5.1. Des méthodes efficaces pour indexer quelques génomes ou jeux<br>de séquençage . . . . . | 88         |
| 2.5.2. Des méthodes tirant difficilement parti de la redondance<br>des données . . . . .       | 88         |
| 2.6. Bibliographie . . . . .   | 89         |
| <b>Chapitre 3. Alignement de séquences . . . . .</b>   | <b>93</b>  |
| Laurent NOÉ  |            |
| 3.1. Introduction . . . . .  | 93         |
| 3.1.1. L'alignement par paire . . . . .  | 93         |
| 3.1.2. Comment évaluer un alignement ? . . . . .   | 94         |
| 3.2. Alignement exact . . . . .  | 96         |
| 3.2.1. Représentation sous forme de graphe d'édition . . . . .                                 | 96         |
| 3.2.2. Alignement global et algorithme de Needleman-Wunch . . . . .                            | 100        |
| 3.2.3. Alignement local et algorithme de Smith-Waterman . . . . .                              | 100        |
| 3.2.4. Alignement avec fonction d'indel affine et algorithme de Gotoh . . . . .                | 102        |
| 3.3. Alignement heuristique . . . . .  | 104        |
| 3.3.1. Graines . . . . .   | 104        |
| 3.3.2. <i>Min-hash</i> et échantillonnage global . . . . .                                     | 110        |
| 3.3.3. <i>Minimizer</i> et échantillonnage local . . . . .                                     | 111        |
| 3.4. Bibliographie . . . . .   | 113        |
| <b>Chapitre 4. Assemblage des génomes . . . . .</b>  | <b>117</b> |
| Dominique LAVENIER   |            |
| 4.1. Introduction . . . . .  | 117        |
| 4.2. Technologies de séquençage . . . . .  | 120        |
| 4.2.1. Lectures courtes . . . . .  | 120        |
| 4.2.2. Lectures longues . . . . .  | 121        |
| 4.2.3. Lectures liées . . . . .  | 122        |
| 4.2.4. Lectures Hi-C . . . . .   | 122        |
| 4.2.5. Cartographie optique . . . . .  | 123        |
| 4.3. Stratégies d'assemblage . . . . .   | 123        |
| 4.3.1. Principales étapes . . . . .  | 123        |
| 4.3.2. Nettoyage et correction des lectures . . . . .  | 124        |
| 4.3.3. Construction des scaffolds . . . . .  | 125        |

|  |     |
|--|-----|
| 4.3.4. Ordonnancement des scaffolds . . . . .            | 127 |
| 4.4. Méthodes de construction des scaffolds . . . . .    | 127 |
| 4.4.1. Assemblage glouton . . . . .                      | 127 |
| 4.4.2. Assemblage OLC . . . . .                          | 129 |
| 4.4.3. Assemblage DBG . . . . .                          | 130 |
| 4.4.4. Assemblage sous contraintes . . . . .             | 133 |
| 4.5. Méthodes d'ordonnancement des scaffolds . . . . .   | 135 |
| 4.5.1. Méthodes basées sur les données Hi-C . . . . .    | 135 |
| 4.5.2. Méthodes basées sur les cartes optiques . . . . . | 140 |
| 4.6. Validation des assemblages . . . . .                | 142 |
| 4.6.1. Métriques . . . . .                               | 142 |
| 4.6.2. Réalignement des lectures . . . . .               | 143 |
| 4.6.3. Prédiction de gènes . . . . .                     | 143 |
| 4.6.4. Compétitions . . . . .                            | 144 |
| 4.7. Conclusion . . . . .                                | 145 |
| 4.8. Bibliographie . . . . .                             | 146 |

## **Chapitre 5. Métagénomique et métatranscriptomique . . . . . 151**

Cervin GUYOMAR et Claire LEMAITRE

|   |     |
|---|-----|
| 5.1. La métagénomique . . . . .   | 151 |
| 5.1.1. Motivations et contexte historique . . . . .   | 151 |
| 5.1.2. Les données métagénomiques . . . . .   | 152 |
| 5.1.3. Défis bioinformatiques pour la métagénomique . . . . .                                   | 155 |
| 5.2. « Qui sont-ils ? » : caractérisation taxonomique<br>des communautés microbiennes . . . . . | 156 |
| 5.2.1. Méthodes pour la métagénomique ciblée . . . . .  | 157 |
| 5.2.2. Méthodes plein-génome avec référence . . . . .   | 158 |
| 5.2.3. Méthodes sans référence . . . . .  | 163 |
| 5.3. « Que font-ils ? » : métagénomique fonctionnelle . . . . .                                 | 169 |
| 5.3.1. Prédiction et annotation de gènes . . . . .  | 169 |
| 5.3.2. Métatranscriptomique . . . . .   | 170 |
| 5.3.3. Reconstruction de réseaux métaboliques . . . . .   | 171 |
| 5.4. Métagénomique comparative . . . . .  | 171 |
| 5.4.1. Métagénomique comparative avec estimation de la diversité . . . . .                      | 172 |
| 5.4.2. Métagénomique comparative <i>de novo</i> . . . . .                                       | 173 |
| 5.5. Conclusion . . . . .   | 177 |
| 5.6. Bibliographie . . . . .  | 178 |

---

|   |     |
|---|-----|
| <b>Chapitre 6. Repliement de l'ARN</b> . . . . .                                    | 187 |
| Yann PONTY et Vladimir REINHARZ   |     |
| 6.1. Introduction . . . . .   | 187 |
| 6.1.1. Repliement des ARN . . . . .   | 188 |
| 6.1.2. La structure secondaire . . . . .  | 190 |
| 6.2. Optimisation pour la prédiction de structure . . . . .                         | 194 |
| 6.2.1. Produire la structure d'énergie minimale . . . . .                           | 194 |
| 6.2.2. Lister les repliements sous-optimaux . . . . .                               | 199 |
| 6.2.3. Repliement comparatif : alignement et repliement conjoint<br>d'ARN . . . . . | 203 |
| 6.3. Approches ensemblistes . . . . .   | 209 |
| 6.3.1. Calcul de la fonction de partition . . . . .                                 | 209 |
| 6.3.2. Échantillonnage statistique . . . . .  | 213 |
| 6.3.3. Probabilité de Boltzmann d'un motif structural . . . . .                     | 218 |
| 6.4. Étudier la structure des ARN en pratique . . . . .                             | 223 |
| 6.4.1. Modèle de Turner . . . . .   | 223 |
| 6.4.2. Outils . . . . .   | 226 |
| 6.5. Bibliographie . . . . .  | 227 |
| <br>  |     |
| <b>Conclusion</b> . . . . .   | 231 |
| Annie CHATEAU et Mikaël SALSON  |     |
| <br>  |     |
| <b>Liste des auteurs</b> . . . . .  | 235 |
| <br>  |     |
| <b>Index</b> . . . . .  | 237 |