

# Table des matières

<b>Avant-propos</b> . . . . .	1
Christine FROIDEVAUX, Marie-Laure MARTIN-MAGNIETTE et Guillem RIGAILL	
<b>Partie 1. Intégration de connaissances</b> . . . . .	7
<b>Chapitre 1. Entrepôts de données cliniques</b> . . . . .	9
Maxime WACK et Bastien RANCE	
1.1. Introduction aux systèmes d'informations cliniques et aux entrepôts biomédicaux : des entrepôts de données pour quels usages ? . . . . .	9
1.1.1. Histoire des entrepôts . . . . .	10
1.1.2. Usage des entrepôts de données aujourd'hui . . . . .	10
1.2. Challenge : des données très dispersées . . . . .	11
1.3. Entrepôts de données et données cliniques . . . . .	12
1.3.1. Structures d'entrepôts . . . . .	12
1.3.2. Construction et alimentation des entrepôts . . . . .	17
1.3.3. Usages . . . . .	18
1.4. Entrepôts et données omiques : difficultés . . . . .	22
1.4.1. Challenges de la volumétrie et de la structuration des données omiques . . . . .	23
1.4.2. Tentatives de réponses . . . . .	23
1.5. Challenges et perspectives . . . . .	24
1.5.1. Vers des entrepôts généralistes . . . . .	25
1.5.2. Dimension éthique de la constitution et de l'utilisation des entrepôts . . . . .	25
1.5.3. Provenance et reproductibilité . . . . .	26
1.5.4. Qualité des données . . . . .	26
1.5.5. Fédération d'entrepôts et partage de données . . . . .	28
1.6. Bibliographie . . . . .	28

<b>Chapitre 2. Méthodes du Web sémantique pour l'intégration de données en sciences de la vie</b> . . . . .	<b>33</b>
Olivier DAMERON	
2.1. Besoins liés aux données en sciences de la vie . . . . .	34
2.1.1. Bases de données en sciences de la vie . . . . .	34
2.1.2. Besoins . . . . .	35
2.1.3. Approches courantes : InterMine et BioMart . . . . .	38
2.2. Web sémantique . . . . .	39
2.2.1. Techniques . . . . .	40
2.2.2. Mise en œuvre . . . . .	51
2.3. Perspectives . . . . .	52
2.3.1. Faciliter l'appropriation par les utilisateurs . . . . .	52
2.3.2. Faciliter l'appropriation par les programmes : données FAIR . . . . .	53
2.3.3. Requêtes fédérées . . . . .	54
2.4. Conclusion . . . . .	55
2.5. Bibliographie . . . . .	56
<b>Chapitre 3. Workflows d'intégration de données bioinformatiques</b> . . . . .	<b>63</b>
Sarah COHEN-BOULAKIA et Frédéric LEMOINE	
3.1. Introduction . . . . .	63
3.2. Chaînes de traitement de données bioinformatiques : difficultés . . . . .	64
3.2.1. Conception d'une chaîne de traitement . . . . .	65
3.2.2. Exécution et reproductibilité d'une analyse . . . . .	66
3.2.3. Maintenance, partage et réutilisation . . . . .	68
3.3. Solutions apportées par les systèmes de <i>workflows</i> scientifiques . . . . .	69
3.3.1. Principes fondamentaux des systèmes de <i>workflows</i> . . . . .	70
3.3.2. Systèmes de <i>workflows</i> . . . . .	74
3.4. Cas d'utilisation : analyse de données de RNA-seq . . . . .	80
3.4.1. Description de l'étude . . . . .	80
3.4.2. De la chaîne de traitement aux <i>workflows</i> . . . . .	83
3.4.3. Implémentation d'une chaîne par un <i>workflow</i> : conclusion . . . . .	87
3.5. Défis, problèmes ouverts et opportunités de recherche . . . . .	88
3.5.1. Formaliser le développement de <i>workflows</i> . . . . .	89
3.5.2. Découverte et partage de <i>workflows</i> . . . . .	90
3.6. Conclusion . . . . .	92
3.7. Bibliographie . . . . .	93

## Partie 2. Intégration et statistiques . . . . . 99

### Chapitre 4. Sélection de variables dans le modèle linéaire général : application à des approches multiomiques pour étudier la qualité des graines . . . . . 101

Céline LÉVY-LEDUC, Marie PERROT-DOCKÈS, Gwendal CUEFF  
et Loïc RAJJOU

4.1. Introduction . . . . .	102
4.2. Méthodologie . . . . .	105
4.2.1. Estimation de la matrice de covariance $\Sigma_q$ . . . . .	105
4.2.2. Estimation de $\mathcal{B}$ . . . . .	108
4.3. Expériences numériques . . . . .	111
4.3.1. Performances statistiques . . . . .	111
4.3.2. Performances numériques . . . . .	115
4.4. Application à l'étude de la qualité des graines . . . . .	115
4.4.1. Données de métabolomique . . . . .	116
4.4.2. Données de protéomique . . . . .	116
4.5. Conclusion . . . . .	119
4.6. Annexes . . . . .	121
4.6.1. Exemple d'utilisation du package <code>MultiVarSel</code> pour l'analyse des données métabolomiques . . . . .	121
4.6.2. Exemple d'utilisation du package <code>MultiVarSel</code> pour l'analyse des données protéomiques . . . . .	123
4.7. Remerciements . . . . .	126
4.8. Bibliographie . . . . .	126

### Chapitre 5. Compression structurée de l'information génétique et étude d'association pangénomique par modèles additifs . . . . . 129

Florent GUINOT, Marie SZAFRANSKI et Christophe AMBROISE

5.1. Études d'association pangénomique . . . . .	130
5.1.1. Introduction à la cartographie génétique et à l'analyse de liaison . . . . .	130
5.1.2. Principes des études d'association pangénomique . . . . .	131
5.1.3. Polymorphisme d'un seul nucléotide . . . . .	132
5.1.4. Pénétrance de la maladie et <i>odds ratio</i> . . . . .	135
5.1.5. Analyse simple-marqueur . . . . .	136
5.1.6. Analyse multimarqueur . . . . .	139
5.2. Compression structurée et étude d'association . . . . .	145
5.2.1. Contexte . . . . .	145
5.2.2. Nouvelle approche par compression structurée . . . . .	146
5.3. Application à la spondylarthrite ankylosante . . . . .	155

5.3.1. Données . . . . .	155
5.3.2. Étude de puissance . . . . .	156
5.3.3. Diagramme de Manhattan . . . . .	157
5.3.4. Estimation pour les agrégats de SNP les plus significatifs . . . . .	157
5.4. Conclusion . . . . .	158
5.5. Bibliographie . . . . .	160

## **Chapitre 6. Des noyaux pour les omiques . . . . . 165**

Jérôme MARIETTE et Nathalie VIALANEIX

6.1. Introduction . . . . .	166
6.2. Données relationnelles . . . . .	167
6.2.1. Données décrites par un noyau . . . . .	167
6.2.2. Données décrites par une mesure de (dis)similarité générale . . . . .	169
6.3. Analyse exploratoire pour des données relationnelles . . . . .	172
6.3.1. Classification non supervisée à noyau . . . . .	173
6.3.2. Analyse en composantes principales à noyau . . . . .	176
6.3.3. Cartes auto-organisatrices à noyau . . . . .	178
6.3.4. Limites des approches relationnelles en apprentissage . . . . .	181
6.4. Combiner les données relationnelles . . . . .	183
6.4.1. Intégration de données en biologie des systèmes . . . . .	183
6.4.2. Place des approches à noyaux dans l'intégration de données . . . . .	184
6.4.3. Un noyau consensuel . . . . .	187
6.4.4. Un noyau parcimonieux qui préserve la topologie des données initiales . . . . .	188
6.4.5. Un noyau complet préservant la topologie des données de départ . . . . .	190
6.5. Application . . . . .	191
6.5.1. Chargement des données Tara Océan . . . . .	192
6.5.2. Intégration des données par approches à noyaux . . . . .	192
6.5.3. Analyse exploratoire : ACP à noyau . . . . .	194
6.6. Information de session pour les résultats de l'exemple . . . . .	201
6.7. Bibliographie . . . . .	203

## **Chapitre 7. Modèles multivariés pour l'intégration de données et la sélection de biomarqueurs dans les données omiques . . . . . 211**

Sébastien DÉJEAN et Kim-Anh LÊ CAO

7.1. Introduction . . . . .	211
7.2. Contexte . . . . .	214
7.2.1. Notations mathématiques . . . . .	214

7.2.2. Terminologie . . . . .	214
7.2.3. Approches multivariées basées sur la projection . . . . .	215
7.2.4. Critère de maximisation spécifique à chaque méthodologie . . . . .	215
7.2.5. Combinaison linéaire de variables pour réduire la taille des données . . . . .	216
7.2.6. Identifier un sous-ensemble de caractéristiques moléculaires pertinentes . . . . .	216
7.2.7. Résumé . . . . .	217
7.3. De la question biologique à l'analyse statistique . . . . .	217
7.3.1. Exploration d'un jeu de données : l'analyse en composantes principales . . . . .	218
7.3.2. Échantillons classifié : projection dans l'analyse discriminante de structure latente . . . . .	224
7.3.3. Intégration de deux ensembles de données : régression linéaire des moindres carrés et méthodes liées . . . . .	228
7.3.4. Intégration de plusieurs ensembles de données : approches multiblocs . . . . .	233
7.4. Résultats graphiques . . . . .	236
7.4.1. Graphiques individuels . . . . .	236
7.4.2. Graphes de variables . . . . .	239
7.5. Résumé général . . . . .	241
7.6. Étude de toxicité hépatique . . . . .	241
7.6.1. Ensembles de données . . . . .	241
7.6.2. Questions biologiques et méthodes statistiques . . . . .	242
7.6.3. Analyse d'un seul jeu de données . . . . .	242
7.6.4. Analyse intégrative . . . . .	247
7.7. Conclusion . . . . .	256
7.8. Remerciements . . . . .	256
7.9. Annexe Code R reproductible . . . . .	256
7.9.1. Exemples de jeux . . . . .	256
7.9.2. Toxicité du foie . . . . .	261
7.10. Bibliographie . . . . .	265

<b>Liste des auteurs . . . . .</b>	<b>271</b>
------------------------------------	------------

<b>Index . . . . .</b>	<b>275</b>
------------------------	------------