

Introduction à Perl pour les biologistes

Édition
française



Introduction à Perl pour la bioinformatique



O'REILLY®

James Tisdall

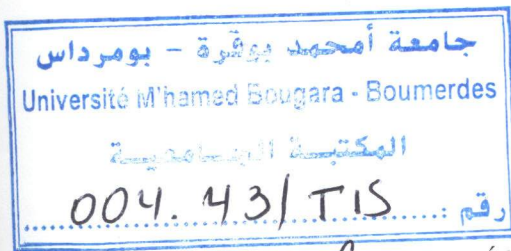
Traduction de Laurent Mouchard et Guénola Ricard

James D. TISDALL



Introduction à Perl pour la bioinformatique

Traduction de
Laurent MOUCHARD et Guénola RICARD



lexp.



Éditions O'REILLY
18 rue Séguier
75006 Paris
FRANCE
france@oreilly.com
<http://www.oreilly.fr>

O'REILLY®

Cambridge • Cologne • Farnham • Paris • Pékin • Sebastopol • Taipei • Tokyo

Table des matières

Préface	ix
1. Biologie et informatique	1
Organisation de l'ADN.....	2
Organisation des protéines	3
In Silico.....	3
Limites du calcul.....	4
2. Premiers pas en Perl	5
Une courbe d'apprentissage basse et longue	5
Avantages de Perl.....	7
Installation de Perl sur votre ordinateur	9
Comment exécuter des programmes ?	12
Éditeurs de texte	14
Obtenir de l'aide	15
3. L'art de la programmation	17
Différentes approches pour la programmation	17
Édition-exécution-modification (et sauvegarde)	18
Sauvegardes	18
Messages d'erreur.....	19
Débogage.....	19
Un vivier de programmes	20
Les programmes libres (open source)	20
Stratégies de programmation.....	20
Le processus de la programmation.....	22

4. Séquences et chaînes de caractères	29
Représentation de données de séquences	29
Un programme pour stocker une séquence d'ADN	32
Concaténation de fragments d'ADN	36
Transcription : ADN en ARN	39
Utilisation de la documentation Perl	42
Calcul du complément inverse d'un brin d'ADN en Perl	43
Protéines, fichiers et tableaux	46
Contexte scalaire et contexte de liste	54
Exercices	56
 5. Motifs et boucles	57
Contrôle du flux du programme	57
Format du code	63
Recherche de motifs	65
Comptage des nucléotides	73
Séparation d'une chaîne de caractères en un tableau	74
Manipulation de chaînes de caractères	81
Écriture dans des fichiers	85
Exercices	89
 6. Sous-programmes et débogage	91
Sous-programmes	91
Portée et sous-programmes	94
Arguments de la ligne de commande et tableaux	100
Transmission de données aux sous-programmes	103
Modules et bibliothèques de sous-programmes	107
Correction des bogues	108
Exercices	121
 7. Mutations et nombres aléatoires	123
Générateurs de nombres aléatoires	124
Un programme qui utilise les nombres aléatoires	125
Un programme pour simuler les mutations dans les séquences d'ADN	131
Générer une séquence d'ADN aléatoire	142
Analyse d'une séquence d'ADN	147
Exercices	154
 8. Le code génétique	157
Hachages	157
Structures de données et algorithmes pour la biologie	158
Le code génétique	163
Traduire de l'ADN en protéines	171
Lire des séquences d'ADN contenues dans des fichiers au format FASTA	175

Cadres de lecture.....	184
Exercices	189
9. Cartes de restriction et expressions régulières	191
Les expressions régulières.....	191
Opérations Perl.....	207
Exercices	207
10. GenBank.....	209
Les fichiers GenBank.....	210
Séparer la séquence des annotations.....	215
Filtrer les annotations	223
Indexer GenBank avec DBM	246
Exercices	250
11. Protein Data Bank.....	253
Vue d'ensemble de la PDB.....	255
Fichiers et répertoires.....	255
Fichiers de la PDB.....	264
Filtrer les fichiers de la PDB	274
Contrôler d'autres programmes.....	285
Exercices	291
12. BLAST.....	293
Obtenir BLAST	293
Recherche de motifs et homologie.....	294
Les fichiers de sortie de BLAST.....	295
Filtrer les sorties BLAST.....	300
Présenter les données	310
Bioperl	314
Exercices	319
13. Aspects avancés du langage Perl.....	321
L'art de la programmation.....	321
Programmation web.....	321
Algorithmes et alignement de séquences	322
Programmation orientée objet.....	322
Modules Perl.....	322
Structures de données complexes.....	323
Bases de données relationnelles	323
Puces à ADN et XML.....	323
Interfaces graphiques.....	324
Modéliser les réseaux.....	324
Ordinateurs à ADN.....	324

A. Ressources	325
Perl.....	325
Informatique.....	328
Bioinformatique	329
Biologie Moléculaire	331
 B. Résumé Perl	333
Interprétation des commandes.....	333
Commentaires.....	333
Valeurs scalaires et variables scalaires.....	334
Affectation.....	335
Instructions et blocs	336
Tableaux.....	337
Hachages.....	337
Opérateurs.....	339
Boucles.....	345
Entrées/Sorties.....	348
Expressions régulières.....	351
Contexte de liste et contexte scalaire.....	357
Sous-programmes et modules.....	358
Fonctions prédéfinies	361
 Index	367

	Débutant	Expérimenté	Spécialiste
Se former			
Agir vite			
Aide-mémoire			

Introduction à Perl pour la bioinformatique



Les dernières avancées majeures en biologie, tels le succès récent du « projet génome humain » (HGP) ou l'accumulation de séquences génomiques complètes, témoignent de la contribution majeure de l'informatique à l'évolution de la biologie, comme ce fut hier le cas pour la physique ou l'astronomie. Avec l'essor de la bioinformatique, les biologistes souhaitent de plus en plus ajouter la programmation à leurs compétences.

Ce livre est une introduction pratique à la programmation en Perl pour les biologistes. Perl excelle dans les fonctions indispensables à cette discipline : traitement de chaînes de caractères, mise en réseau, possibilité de piloter d'autres programmes et prototypage rapide. Cet ouvrage ne nécessite aucune connaissance préalable de la programmation. Il envisage la programmation comme un outil à part entière et désormais indispensable de la recherche en biologie. Dans les premiers chapitres, les principaux concepts sont soigneusement expliqués afin de donner au lecteur une base solide. La programmation est ensuite abordée du point de vue de la biologie et quelques programmes répondant aux principaux besoins d'un laboratoire sont décortiqués. Cette méthode permet de fournir d'une part des programmes clé en main et d'autre part les moyens d'écrire des applications plus spécifiques ou d'adapter un programme existant. Au fur et à mesure du livre les notions de programmation et les problèmes abordés se complexifient. Des exercices d'application sont proposés en fin de chapitre. Après avoir lu ce livre, le lecteur aura des bases solides en Perl et disposera d'un ensemble de programmes concrets comme l'analyse de données provenant de GenBank ou issues de BLAST.

Les thèmes suivants sont abordés :

- concepts fondamentaux de la programmation en Perl ;
- manipulation des séquences d'ADN et des chaînes de caractères ;
- modularisation et débogage du code ;
- reconnaissance de motifs à l'aide d'expressions régulières ;
- simulation de mutations génétiques à l'aide de générateurs de nombres aléatoires ;
- gestion des fichiers au format FASTA ;
- cartes de restriction et expressions régulières ;
- analyse et annotation de séquences ou de structures provenant de banques de données biologiques (GenBank, PDB) ou de résultats fournis par BLAST.

Cet ouvrage servira de tremplin aux biologistes qui, néophytes en programmation, pourront rapidement progresser et réussir leurs projets. Les informaticiens désireux de s'initier à la bioinformatique y trouveront un aperçu de la problématique de cette discipline.

ISBN 2-84177-206-3



9 782841 772063