

Chapter 4

Computational Modeling and Machine Learning in Bioinformatics

Diego Mariano

 <https://orcid.org/0000-0002-5899-2052>

Universidade Federal de Minas Gerais, Brazil

Lucas Moraes dos Santos

Universidade Federal de Minas Gerais, Brazil

Raquel Cardoso de Melo-Minardi

Universidade Federal de Minas Gerais, Brazil

ABSTRACT

In recent years, machine learning has revolutionized the world. Indeed, bioinformatics has benefited from machine learning techniques to make new scientific discoveries, such as producing new biotechnological products, discovering new drugs, and understanding the mechanism of action of diseases, among others. In this chapter, you will learn about machine learning methodologies, the types of machine learning, and the traditional algorithms used to build models. You will also learn about some strategies used for the computational modeling of biological problems with a special focus on the representation of structures in the data pre-processing stage.

DOI: 10.4018/979-8-3693-3192-7.ch004

Copyright © 2025, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

1. INTRODUCTION

Artificial intelligence has brought profound changes to society, especially in life sciences. New high-performance technologies like Next-Generation Sequencing (NGS) and Cryo-electron microscopy (cryo-EM) equipment have increased biological data collection. In this context, machine learning (ML) techniques have allowed knowledge extraction, leading to new scientific discoveries (Dos Santos; Mariano; De Melo-Minardi, 2024). A popular example is the Alphafold algorithm, which uses machine learning to solve the protein folding problem (Jumper et al., 2021).

We like to think that machine learning is the art of predicting the future based on observations made in the past or even deeper observations made in the data itself. Machine learning focuses on extracting knowledge by applying statistical methods to extensive data sets collected from previous observations (Mitchell, 1997).

In this case, it is common for machine learning pipelines to have a feature engineering step (*e.g.*, dimensionality reduction, kernel methods, etc.) or feature selection before applying the learning algorithm. This approach makes it possible to build models capable of making high-performance future predictions (Dos Santos; Mariano; De Melo-Minardi, 2024).

Additionally, ML-based techniques have great applicability in biological sciences, from discovering the mechanism of action of diseases to the production of new drugs (Patel et al., 2020). However, before we talk about the applicability of these algorithms, you need to know the types of machine learning.

1.1 Types of Machine Learning

Traditionally, machine learning methods are divided into supervised, unsupervised, semi-supervised, and reinforcement learning (Duda; Hart; Stork, 2012).

In supervised learning, labels are used in the training and testing stages (Cunningham; Cord; Delany, 2008). Thus, models are built to detect patterns based on the answers they already have. Examples of supervised learning techniques are classification and regression. In classification, the model tries to predict groups, while in regression, the model tries to predict discrete values.

Meanwhile, in unsupervised learning, only features are used in pattern detection. Examples of unsupervised machine learning techniques are clustering and dimensionality reduction.

Semi-supervised learning methodologies bring together characteristics of supervised and unsupervised learning. This technique is generally used when little labeled data is available (Zhu, 2005).

28 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/computational-modeling-and-machine-learning-in-bioinformatics/361320

Related Content

Improving PSI-BLAST's Fold Recognition Performance through Combining Consensus Sequences and Support Vector Machine

Ren-Xiang Yan, Jing Liu and Yi-Min Tao (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications* (pp. 1667-1675).

www.irma-international.org/chapter/improving-psi-blast-fold-recognition/76140

Chaos Game Representation of Mitochondrial Genomes: Markov Chain Model Simulation and Vertebrate Phylogeny

Zu-Guo Yu, Guo-Sheng Han, Bo Li, Vo Anhand Yi-Quan Li (2011). *Interdisciplinary Research and Applications in Bioinformatics, Computational Biology, and Environmental Sciences* (pp. 28-38).

www.irma-international.org/chapter/chaos-game-representation-mitochondrial-genomes/48362

Medical Survival Analysis Through Transduction of Semi-Supervised Regression Targets

Faisal M. Khan and Qihua Liu (2011). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 52-65).

www.irma-international.org/article/medical-survival-analysis-through-transduction/63617

Efficient and Robust Analysis of Large Phylogenetic Datasets

Sven Rahmann, Tobias Muller, Thomas Dandekar and Matthias Wolf (2006). *Advanced Data Mining Technologies in Bioinformatics* (pp. 104-117).

www.irma-international.org/chapter/efficient-robust-analysis-large-phylogenetic/4248

Transcriptome-To-Metabolome™ Biosimulation Reveals Human Hippocampal Hypometabolism with Age and Alzheimer's Disease

Clyde F. Phelix, Richard G. LeBaron, Dawnlee J. Roberson, Rosa E. Villanueva, Greg Villareal, Omid B. Rahimi, Sandra Siedlak, Xiongwei Zhu and George Perry (2011). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 1-18).

www.irma-international.org/article/transcriptome-metabolome-biosimulation-reveals-human/62298