जैव सूचना विज्ञान क्षेत्र में डाटा माइनिंग के अनुप्रयोगः एक समीक्षा

शबाना बेगम[1], सुभ्रजित सत्पथी[2], समर्थ गोदारा[3], सपना निगम[3], अक्षय धीरज[4],
ई. युवराज[1], मो. अशराफुल हक[3], संचिता नाहा[3]

10.18805/BKAP428

सारांश

अगली पीढ़ी की अनुक्रमण तकनीक (NGS) प्रचुर मात्रा में जीनोमिक जानकारी प्रदान करती है। जैव सूचना विज्ञान का मुख्य कार्य जीनोमिक डेटा के सभी पहलुओं को समझना एवं समस्यायों का समाधान करना बन गया है। यह लेख मुख्य रूप से जैव सूचना विज्ञान में डेटा माइनिंग के अनुप्रयोगो का विश्लेषण करता है। जैव सूचना विज्ञान की परिभाषा और मुख्य शोध सामग्री के सारांश के आधार पर, जैव सूचना विज्ञान डेटा के प्रसंस्करण प्रवाह को चित्रित किया गया है। फिर जैव सूचना विज्ञान में डेटा प्रीप्रोसेसिंग, आयाम में कमी और सांख्यिकीय मशीन लर्निंग के परिप्रेक्ष्य से डेटा खनन का प्रभावी परिचय देता है। जैव सूचना विज्ञान के क्षेत्र में डेटा माइनिंग के अनुप्रयोग को समझाया गया है। यह जैव सूचना विज्ञान में डेटा माइनिंग की कुछ मौजूदा चुनौतियों और अवसरों पर भी प्रकाश डालता है।

शब्द कुंजी: जैव सूचना विज्ञान उपकरण, जैव सूचना विज्ञान, डेटा माइनिंग, प्रोटीन अनुक्रम विश्लेषण।

Applications of Data Mining in the Bioinformatics Field: A Review

Shbana Begam[1], Subhrajit Satpathy[2], Samarth Godara[3], Sapna Nigam[3],
Akshay Dheeraj[4], I. Yuvaraj[1], Md. Ashraful Haque[3], Sanchita Naha[3]

ABSTRACT

The next generation sequencing (NGS) technology generates a large amount of genomic data. Bioinformatics has made it a priority to deal with genomic data in all of its forms. The main focus of this article is on the use of data mining in bioinformatics. The processing flow of bioinformatics data is depicted using a summary of the definition and relevant research contents of bioinformatics. Then emphatically introduces data mining from the perspective of data preprocessing, dimension reduction and statistical machine learning in bioinformatics. The use of data mining in the field of bioinformatics is discussed. It also discusses some of the current obstacles and prospects in bioinformatics data mining.

Key words: Bioinformatics tools, Bioinformatics, Data mining, Protein sequences analysis.

हाल के वर्षों में, जीनोमिक्स और प्रोटिओमिक्स में तेजी से विकास ने बड़ी मात्रा में जैविक डेटा उत्पन्न किया है। इन आंकड़ों से निष्कर्ष निकालने के लिए परिष्कृत कम्प्यूटेशनल विश्लेषण की आवश्यकता होती है। जैव सूचना विज्ञान, एक प्रकार का कम्प्यूटेशनल जीव विज्ञान, सूचना प्रौद्योगिकी और कंप्यूटर विज्ञान का उपयोग करके जैविक डेटा की व्याख्या करने का अंतः विषय विज्ञान है। क्योंकि हम बड़ी मात्रा में जीनोमिक, प्रोटिओमिक और अन्य डेटा उत्पन्न और एकीकृत करना जारी रखेंगे अंतत जांच के इस नए क्षेत्र का महत्व आगे और बढ़ेगा। जैव सूचना विज्ञान में अनुसंधान का एक विशेष रूप से सक्रिय क्षेत्र जैविक समस्याओं को हल करने के लिए डेटा माइनिंग तकनीकों का अनुप्रयोग और विकास है (Yuan, 2016)। बड़े जैविक डेटा सेट का विश्लेषण करने के लिए डेटा

[1]ICAR-National Institute for Plant Biotechnology, New Delhi-110 012, India.
[2]International Crops Research Institute for the Semi-Arid Tropics, Hyderabad-502 324, Telangana, India.
[3]ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110 012, India.
[4]ICAR-Indian Institute of Soil and Water Conservation, Dehradun-248 195, Uttarakhand, India.

Corresponding Author: Shbana Begam, ICAR-National Institute for Plant Biotechnology, New Delhi-110 012, India.
Email: shaba.shb@gmail.com

How to cite this article: Begam, S., Satpathy, S., Godara, S., Nigam, S., Dheeraj, A., Yuvaraj, I., Haque, M.A. and Naha, S. (2022). Applications of Data Mining in the Bioinformatics Field: A Review. Bhartiya Krishi Anusandhan Patrika. 37(2): 121-125. DOI: 10.18805/BKAP428.

Submitted: 10-01-2022 **Accepted:** 28-05-2022 **Online:** 20-06-2022

से संरचना या सामान्यीकरण का अनुमान लगाकर डेटा की समझ बनये रखने की आवश्यकता होती है। इस प्रकार के विश्लेषण के उदाहरणों में प्रोटीन संरचना भविष्यवाणी, जीन वर्गीकरण, माइक्रोएरे डेटा के आधार पर कैंसर वर्गीकरण, जीन अभिव्यक्ति डेटा की क्लस्टरिंग, प्रोटीन–प्रोटीन इंटरैक्शन के सांख्यिकीय मॉडलिंग आदि शामिल हैं। इसलिए, हम डेटा के बीच बातचीत को बढ़ाने की एक बड़ी क्षमता देखते हैं, जो की खनन और जैव सूचना विज्ञान है।

डेटा माइनिंग

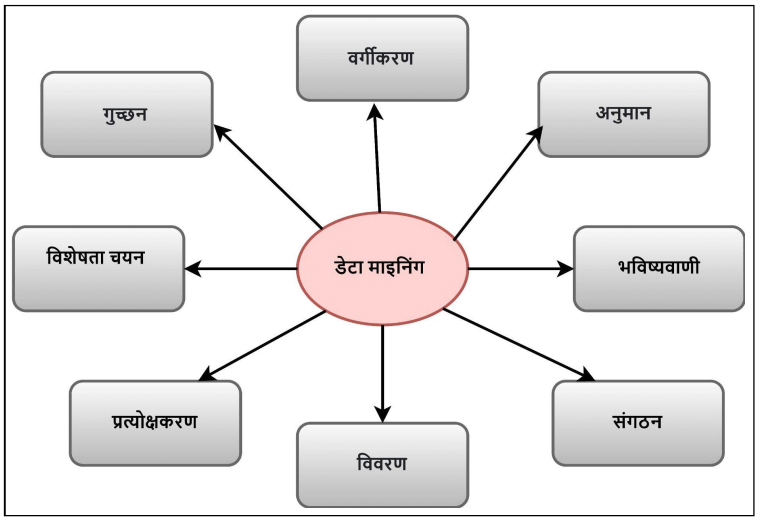
डेटा माइनिंग का तात्पर्य बड़ी मात्रा में उपलब्ध डेटा से ज्ञान निकालने या ''खनन'' करने से है। डेटा माइनिंग (DM) बड़ी मात्रा में डेटा में नए रोचक पैटर्न और संबंध खोजने का विज्ञान है। इसे वेयरहाउसों में संग्रहीत डेटा की बड़ी मात्रा में खुदाई करके सार्थक नए सहसंबंध, पैटर्न और प्रवृत्तियों की खोज की प्रक्रिया के रूप में परिभाषित किया गया है। डेटा माइनिंग को कभी–कभी डेटाबेस में नॉलेज डिस्कवरी (KDD) भी कहा जाता है। डेटा माइनिंग किसी उद्योग के लिए विशिष्ट नहीं है, इसके लिए बुद्धिमान प्रौद्योगिकियों और डेटा में छिपे हुए, ज्ञान की संभावना का पता लगाने की आवश्यकता होती है। डेटा माइनिग दृष्टिकाण जैव सूचना विज्ञान के लिए आदर्श रूप से अनुकूल लगता है, क्योंकि यह डेटा–समृद्ध है, लेकिन आणविक स्तर पर जीवन के संगठन के व्यापक सिद्धांत का अभाव है। जैविक जानकारी के व्यापक डेटाबेस उपन्यास KDD विधियों के विकास के लिए चुनौतियां और अवसर दोनों पैदा करते हैं। खनन जैविक डेटा जीव विज्ञान में एकत्र किए गए बड़े पैमाने पर डेटासेट से उपयागी ज्ञान निकालने में मदद करता है और अन्य संबंधित जीवन विज्ञान क्षेत्रों जैसे कि चिकित्सा और तंत्रिका विज्ञान में।

डेटा माइनिंग के कार्य

डेटा माइनिग के दो ''उच्च–स्तरीय'' प्राथमिक लक्ष्य, भविष्यवाणी और विवरण हैं। डेटा माइनिंग के लिए उपयुक्त मुख्य कार्य चित्र 1 में दर्शित है, जिनमें से सभी में डेटा से सार्थक नए पैटर्न का खनन शामिल है:

- **वर्गीकरण:** वर्गीकरण डेटा आइटम को कई पूर्वनिर्धारित वर्गों में से एक में मैप (वर्गीकृत) करता है।
- **अनुमान:** कुछ इनपुट डेटा को देखते हुए, कुछ अज्ञात निरंतर चर के लिए एक परिणाम के साथ आना।
- **भविष्यवाणी:** वर्गीकरण और अनुमान के समान ही, अभिलेखों को कुछ भविष्य के व्यवहार या अनुमानित भविष्य के मूल्य के अनुसार वर्गीकृत किया जाता है।
- **संगठन नियम:** यह निर्धारित करना कि कौन सी चीजें एक साथ चलती हैं, इसे निर्भरता मॉडलिंग भी कहा जाता है।
- **गुच्छन:** जनसंख्या को कई उपसमूहों या समूहों में विभाजित करना।
- **विवरण और प्रत्याक्षकरण:** विजुअलाइजेशन तकनीकों का उपयाग करके डेटा का प्रतिनिधित्व करना।

डेटा का अधिगम निम्नलिखित दो श्रेणियों में होता है: निर्देशित (''पर्यवेक्षित'') और अप्रत्यक्ष (''अनपर्यवेक्षित'') अधिगम। पहले तीन कार्ये–वर्गीकरण, अनुमान और भविष्यवाणी–पर्यवेक्षित शिक्षण के उदाहरण हैं। अगले तीन कार्ये–संगठन नियम, गुच्छन और विवरण और प्रत्योक्षकरण–अप्रशिक्षित शिक्षण के उदाहरण हैं। अनुपयोगी शिक्षण में, लक्ष्य के रूप में किसी भी



चित्र 1: डेटा माइनिंग के कार्य।

चर का चयन नहीं किया जाता है लक्ष्य सभी चरों के बीच कुछ संबंध स्थापित करना है। बिना पर्यवेक्षित शिक्षण किसी विशेष लक्ष्य क्षेत्र के उपयोग के बिना पैटर्न खोजने का प्रयास करता है। नए डेटा माइनिंग और नॉलेज डिस्कवरी टूल्स का विकास सक्रिय शोध का विषय है। इन उपकरणों के विकास के पीछे एक प्रेरणा आधुनिक जीव विज्ञान में उनका संभावित अनुप्रयोग है (Diniz and Canduri, 2017)।

जैवसूचना विज्ञान

जैवसूचना विज्ञान शब्द को 1979 में पॉलिन हॉगवेग द्वारा जैविक प्रणालियों में सूचनात्मक प्रक्रियाओं के अध्ययन के लिए गढ़ा गया था। 1980 के दशक के उत्तरार्ध से इसका प्राथमिक उपयोग जीनोमिक्स और आनुवंशिकी में किया गया, विशेष रूप से जीनोमिक्स के उन क्षेत्रों में जिसमें बड़े पैमाने पर डीएनए अनुक्रमण शामिल है। जैव सूचना विज्ञान को जैविक सूचना के प्रबंधन के लिए कंप्यूटर प्रौद्योगिकी के अनुप्रयोग के रूप में परिभाषित किया जा सकता है (Lesk, 2019)।

जैव सूचना विज्ञान जैविक अनुक्रमों और अणुओं से जानकारी को संग्रहीत करने, निकालने, व्यवस्थित करने, विश्लेषण करने, व्याख्या करने और उपयोग करने का विज्ञान है। यह मुख्य रूप से डीएनए अनुक्रमण और मानचित्रण तकनीकों में प्रगति से प्रेरित है। पिछले कुछ दशकों में जीनोमिक और अन्य आणविक अनुसंधान प्रौद्योगिकियों में तेजी से विकास और सूचना प्रौद्योगिकियों के विकास ने आणविक जीव विज्ञान से संबंधित भारी मात्रा में जानकारी का उत्पादन किया है (Fulekar, 2009)। जैव सूचना विज्ञान का प्राथमिक लक्ष्य जैविक प्रक्रियाओं की समझ को बढ़ाना है। जैव सूचना विज्ञान में अनुसंधान के कुछ प्रमुख क्षेत्रों में शामिल है।

अनुक्रम विश्लेषण

अनुक्रम विश्लेषण, कम्प्यूटेशनल जीव विज्ञान में सबसे आदिम कार्यवाही है। इस कार्यवाही में यह पता लगाना शामिल है कि चिकित्सा विश्लेषण और जीनोम मैपिंग प्रक्रियाओं के दौरान जैविक अनुक्रमों का कौन सा हिस्सा समान है और कौन सा हिस्सा भिन्न है। अनुक्रम विश्लेषण का तात्पर्य कंप्यूटर पर अनुक्रम संरेखण, अनुक्रम डेटाबेस, बार–बार अनुक्रम खोज, या अन्य जैव सूचना विज्ञान विधियों के लिए डीएनए या पेप्टाइड अनुक्रम के अधीन है।

जीनोम एनोटेशन

जीनोमिक्स के संदर्भ में, एनोटेशन एक डीएनए अनुक्रम में जीन और अन्य जैविक विशेषताओं को चिह्नित करने की प्रक्रिया है। पहला जीनोम एनोटेशन सॉफ्टवेयर सिस्टम 1995 में डॉ. ओवेन व्हाइट द्वारा डिजाइन किया गया था।

जीन अभिव्यक्ति का विश्लेषण

कई जीनों की अभिव्यक्ति को विभिन्न तकनीकों जैसे कि माइक्रोएरे, एक्सप्रेस्ड सीडीएनए अनुक्रम टैग अनुक्रमण, जीन अभिव्यक्ति के सीरियल विश्लेषण टैग अनुक्रमण, बड़े पैमाने पर समानांतर हस्ताक्षर अनुक्रमण, या विभिन्न अनुप्रयोगों के साथ एमआरएनए स्तरों को मापने के द्वारा निर्धारित किया जा सकता है। बहुसंकेतन इन–सीटू संकरण आदि, ये सभी तकनीकें अत्यधिक शोर–प्रवण हैं और जैविक माप में पूर्वाग्रह के अधीन हैं। यहां प्रमुख अनुसंधान क्षेत्र में उच्च–थ्रूपुट जीन अभिव्यक्ति अध्ययनों में शोर से संकेत को अलग करने के लिए सांख्यिकीय उपकरण विकसित करना शामिल है।

प्रोटीन अभिव्यक्ति का विश्लेषण

जीन अभिव्यक्ति को एमआरएनए और प्रोटीन अभिव्यक्ति सहित कई तरीकों से मापा जाता है, हालांकि प्रोटीन अभिव्यक्ति वास्तविक जीन गतिविधि के सर्वोत्तम संकेतों में से एक है क्योंकि प्रोटीन आमतौर पर सेल गतिविधि के अंतिम उत्प्रेरक होते हैं। प्रोटीन माइक्राएरे और हाई–थ्रूपुट मास स्पेक्ट्रोमेट्री एक जैविक नमूने में मौजूद प्रोटीन का एक स्नैपशॉट प्रदान कर सकते हैं। जैव सूचना विज्ञान प्रोटीन माइक्राएरे और एचटी एमएस डेटा को समझने में बहुत अधिक शामिल है।

प्रोटीन संरचना भविष्यवाणी

एक प्रोटीन (तथाकथित, प्राथमिक संरचना) के अमीनो एसिड अनुक्रम को जीन पर अनुक्रम से आसानी से निर्धारित किया जा सकता है जो इसके लिए कोड करता है। ज्यादातर मामलों में, यह प्राथमिक संरचना विशिष्ट रूप से अपने मूल वातावरण में एक संरचना निर्धारित करती है। प्रोटीन के कार्य को समझने के लिए इस संरचना का ज्ञान महत्वपूर्ण है। बेहतर शब्दों की कमी के कारण, संरचनात्मक जानकारी को आमतौर पर द्वितीयक, तृतीयक और चतुर्धातुक संरचना के रूप में वर्गीकृत किया जाता है। प्रोटीन संरचना भविष्यवाणी दवा डिजाइन और उपन्यास एंजाइमों के डिजाइन के लिए सबसे महत्वपूर्ण में से एक है। ऐसी भविष्यवाणियों का एक सामान्य समाधान शोधकर्ताओं के लिए एक खुली समस्या बनी हुई है।

तुलनात्मक जीनोमिक्स

तुलनात्मक जीनोमिक्स विभिन्न जैविक प्रजातियों में जीनोम संरचना और कार्य के संबंध का अध्ययन है। जीन की खोज

तुलनात्मक जीनोमिक्स का एक महत्वपूर्ण अनुप्रयोग है, जैसा कि जीनोम के नए, गैरंदककोडिंग कार्यात्मक तत्वों की खोज है। तुलनात्मक जीनोमिक्स विभिन्न जीवों के प्रोटीन, आरएनए और नियामक क्षेत्रों में समानता और अंतर दोनों का फायदा उठाता है। जीनोम तुलना के लिए कम्प्यूटेशनल दृष्टिकोण हाल ही में कंप्यूटर विज्ञान में एक सामान्य शोध विषय बन गया है।

मॉडलिंग जैविक प्रणाली

जैविक प्रणालियों की मॉडलिंग जीव विज्ञान और गणितीय जीव विज्ञान प्रणालियों का एक महत्वपूर्ण कार्य है। कम्प्यूटेशनल सिस्टम बायोलॉजी का उद्देश्य कंप्यूटर मॉडलिंग के लक्ष्य के साथ बड़ी मात्रा में जैविक डेटा के एकीकरण के लिए कुशल एल्गोरिदम, डेटा संरचनाओं, विजुअलाइजेशन और संचार उपकरणों का विकास और उपयोग करना है। इसमें जैविक प्रणालियों के कंप्यूटर सिमुलेशन का उपयोग शामिल है, जैसे सेलुलर सबसिस्टम, मेटाबोलाइट्स और एंजाइमों के नेटवर्क, सिग्नल ट्रांसडक्शन पथ और जीन नियामक नेटवर्क दोनों इन सेलुलर प्रक्रियाओं के जटिल कनेक्शन का विश्लेषण और कल्पना करते हैं। कृत्रिम जीवन सरल जीवन रूपों के कंप्यूटर सिमुलेशन के माध्यम से विकासवादी प्रक्रियाओं को समझने का एक प्रयास है।

उच्च–थ्रूपुट छवि विश्लेषण

कम्प्यूटेशनल तकनीकों का उपयोग बड़ी मात्रा में उच्च सूचना सामग्री बायोमेडिकल छवियों के प्रसंस्करण, परिमणीकरण और विश्लेषण को तेज करने या पूरी तरह से स्वचालित करने के लिए किया जाता है। आधुनिक छवि विश्लेषण प्रणालियां छवियों के एक बड़े या जटिल सेट से माप करने के लिए एक पर्यवेक्षक की क्षमता को बढ़ाती हैं। एक पूरी तरह से विकसित विश्लेषण प्रणाली पूरी तरह से पर्यवक्षक की जगह ले सकती है। निदान और अनुसंधान दोनों के लिए बायोमेडिकल इमेजिंग अधिक महत्वपूर्ण होती जा रही है। इस क्षेत्र में अनुसंधान के कुछ उदाहरण हैं: नैदानिक छवि विश्लेषण और विजुअलाइजेशन, डीएनए मैपिंग, बायोइमेज इंफॉर्मेटिक्स, आदि में क्लोन ओवरलैप्स का उल्लेख करना (Gaur *et al.*, 2021)।

प्रोटीन–प्रोटीन डॉकिंग

पिछले दो दशकों में, एक्स–रे क्रिस्टलोग्राफी और प्रोटीन परमाणु चुंबकीय अनुनाद स्पेक्ट्रोस्कोपी (प्रोटीन एनएमआर) द्वारा हजारों प्रोटीन त्रि–आयामी संरचनाएं निर्धारित की गई हैं। जैविक वैज्ञानिक के लिए एक केंद्रीय प्रश्न यह है कि क्या प्रोटीन–प्रोटीन इंटरैक्शन प्रयोग किए बिना, केवल इन 3 डी आकृतियों के आधार पर संभावित प्रोटीन–प्रोटीन इंटरैक्शन की भविष्यवाणी करना व्यावहारिक है। प्रोटीन–प्रोटीन डॉकिंग समस्या से निपटने के लिए कई तरह के तरीके विकसित किए गए हैं, हालांकि ऐसा लगता है कि इस क्षेत्र में अभी भी बहुत काम किया जाना बाकी है।

जैव सूचना विज्ञान में डाटा माइनिंग का अनुप्रयोग

जैव सूचना विज्ञान में डेटा माइनिंग के अनुप्रयोगों में जीन खोज, प्रोटीन फंक्शन डोमेन डिटेक्शन, फंक्शन मोटिफ डिटेक्शन, प्रोटीन फंक्शन इंट्रेंस, डिजीज डायग्नोसिस, डिजीज प्रोग्नोसिस, डिजीज ट्रीटमेंट ऑप्टिमाइजेशन, प्रोटीन और जीन इंटरेक्शन नेटवर्क रिकंस्ट्रक्शन, डेटा क्लींजिंग और प्रोटीन सब–सेलुलर लोकेशन प्रेडिक्शन शामिल हैं। उदाहरण के लिए, रोगी के परिणाम की भविष्यवाणी करने के लिए माइक्राएरे प्रौद्योगिकियों का उपयाग किया जाता है। मरीजों के जीनोटाइपिक माइक्राएरे डेटा के आधार पर, उनके जीवित रहने का समय और ट्यूमर मेटास्टेसिस या पुनरावृत्ति के जोखिम का अनुमान लगाया जा सकता है। मास स्पेक्ट्रोस्कोपी के माध्यम से पेप्टाइड पहचान के लिए मशीन लर्निंग का उपयाग किया जा सकता है। डेटाबेस खोज द्वारा पेप्टाइड पहचान के लिए स्टोकेस्टिक बेमेल को कम करने में एक अग्रानुक्रम द्रव्यमान स्पेक्ट्रम में टुकड़े आयनों के बीच सहसंबंध महत्वपूर्ण है (Mining, 2011)। एक कुशल स्कोरिंग एल्गोरिथ्म जो एक ट्यून करने योग्य और व्यापक तरीके से सहसंबंधी जानकारी पर विचार करता है, वह अत्यधिक वांछनीय है।

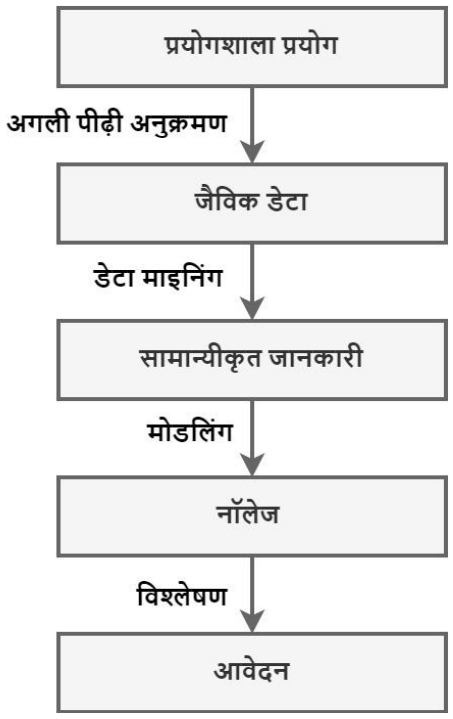
जैव सूचना विज्ञान के क्षत्र में डेटा माइनिंग के कुछ अनुप्रयोगो की सूचि निम्नानुसार हैं:

1. जीन फाइंडिंग।
2. प्रोटीन फंक्शन डोमेन डिटेक्शन।
3. फंक्शन मोटिफ डिटेक्शन।
4. प्रोटीन फंक्शन इंट्रेंस।
5. रोग निदान।
6. रोग निदान।
7. राग उपचार अनुकूलन।
8. प्रोटीन और जीन इंटरैक्शन नेटवर्क।
9. पुनर्निमाण।
10. डेटा सफाई।
11. प्रोटीन उपंदकसेलुलर स्थान भविष्यवाणी।
12. प्रोटीन और डीएनए अनुक्रमों का विश्लेषण।

जैव सूचना विज्ञान डेटा का प्रसंस्करण प्रवाह

बड़ी मात्रा में जैविक डेटा का डेटा खनन और प्रसंस्करण जैव

सूचना विज्ञान का एक महत्वपूर्ण कार्य है। प्रीप्रोसेसिंग, डेटा माइनिंग और बड़े पैमाने पर जैविक डेटा का विश्लेषण और मात्रात्मक या गुणात्मक के गणितीय मॉडल की स्थापना के माध्यम से, शोधकर्ता प्रयोग डेटा, जैविक प्रक्रियाओं और प्रमुख बीमारियों के बीच संबंधित गुणों पर और चर्चा कर सकते हैं। विभिन्न प्रकार के जैविक डेटा जैसे जीनोम–सीक्वेंसिंग टेक्स्ट, जीन अभिव्यक्ति स्तर, सेल फेनोटाइप और संबंधित तंत्र के व्यवस्थित खनन और रोग के मॉडल का एकीकरण रोग की भविष्यवाणी और उपचार के लिए बहुत महत्व रखता है। जैसा कि चित्र 2 में दिखाया गया है, जीव विज्ञान प्रयोग का कच्चा डेटा एकत्र और मानकीकृत होने के बाद सामान्यीकृत डेटा में बदल जाता है, फिर संबंधित गणितीय मॉडल को संभावित तंत्र को जोड़ने का पता लगाने के लिए स्थापित किया जा सकता है, इसलिए भविष्यवाणी और पूर्वानुमान के लिए विश्वसनीय समर्थन प्रदान किया जा सकता है। कुछ रोग विज्ञान में विभिन्न प्रकार की रोग समस्याओं को हल करने के लिए यह सामान्य प्रक्रिया प्रवाह है। हालांकि, कुछ बीमारियों के लिए, अधिक सटीक विश्लेषण परिणाम प्राप्त करने के लिए अभी भी एक विशेष गणना पद्धति की आवश्यकता है (Raza, 2012)।



चित्र 2: जैव सूचना विज्ञान डेटा का प्रसंस्करण प्रवाह।

निष्कर्ष

जैव सूचना विज्ञान और डेटा खनन एक अंतः विषय विज्ञान के रूप में विकसित हो रहे हैं। डेटा माइनिंग दृष्टिकोण जैव सूचना विज्ञान के लिए आदर्श रूप से अनुकूल लगता है क्योंकि जैव सूचना विज्ञान डेटा–समृद्ध है लेकिन आणविक स्तर पर जीवन के संगठन के व्यापक सिद्धांत का अभाव है। हालाँकि, जैव सूचना विज्ञान में डेटा माइनिंग जैविक डेटाबेस के कई पहलुओं से बाधित है, जिसमें उनके आकार, संख्या, विविधता और एक मानक ऑन्कोलॉजी की कमी शामिल है, जो उनकी पूछताछ के साथ–साथ गुणवत्ता और उत्पत्ति की जानकारी के विषम डेटा में सहायता करता है। एक अन्य समस्या संभावित उपयोगकर्ताओं के बीच मौजूद विशेषज्ञता के डोमेन स्तरों की सीमा है, इसलिए डेटाबेस क्यूरेटर के लिए सभी के लिए उपयुक्त पहुंच तंत्र प्रदान करना मुश्किल हो सकता है। जैविक डेटाबेस का एकीकरण भी एक समस्या है। डेटा माइनिंग और जैव सूचना विज्ञान आज तेजी से बढ़ते अनुसंधान क्षेत्र हैं। यह जांचना महत्वपूर्ण है कि जैव सूचना विज्ञान में महत्वपूर्ण शोध मुद्दे क्या हैं और स्केलेबल और प्रभावी विश्लेषण के लिए नई डेटा खनन विधियों को विकसित करना महत्वपूर्ण है।

संदर्भ

Diniz, W.J.D.S. and Canduri, F. (2017). Bioinformatics: An overview and its applications. Genet. Mol. Res. 16(1): 10-4238.

Fulekar, M.H. (Ed.). (2009). Bioinformatics: Applications in Life and Environmental Sciences. Springer Science and Business Media.

Gaur, L., Solanki, A., Wamba, S.F. and Jhanjhi, N.Z. (Eds.). (2021). Advanced AI Techniques and Applications in Bioinformatics. CRC Press.

Mining, D. (2011). What is Data Mining. J. Frand's web page at UCLA page (available from www. Anderson. UCLA. edu/faculty/json. frand/teacher/technologies/palace/ data mining. htm)–visited on, 29(05).

Raza, K. (2012). Application of Data Mining in Bioinformatics. arXiv Preprint arXiv:1205.1125.

Lesk, A. (2019). Introduction to Bioinformatics. Oxford University Press.

Yuan, M. (2016). Study on the Application of Data Mining in Bioinformatics. Advances in Engineering Research. https://doi.org/10.2991/icmeit-16.2016.21.