Biological Data

Subjects: Others
Contributor: HandWiki Xu

Biological data refers to a compound or information derived from living organisms and their products. A medicinal compound made from living organisms, such as a serum or a vaccine, could be characterized as biological data. Biological data is highly complex when compared with other forms of data. There are many forms of biological data, including text, sequence data, protein structure, genomic data and amino acids, and links among others.

Keywords: medicinal compound ; living organisms ; vaccine

1. Biological Data and Bioinformatics

Biological data works closely with Bioinformatics, which is a recent discipline focusing on addressing the need to analyze and interpret vast amounts of genomic data.

In the past few decades, leaps in genomic research have led to massive amounts of biological data. As a result, bioinformatics was created as the convergence of genomics, biotechnology, and information technology, while concentrating on biological data.

Biological Data has also been difficult to define, as bioinformatics is a wide-encompassing field. Further, the question of what constitutes as being a living organism has been contentious, as "alive" represents a nebulous term that encompasses molecular evolution, biological modeling, biophysics, and systems biology. From the past decade onwards, bioinformatics and the analysis of biological data have been thriving as a result of leaps in technology required to manage and interpret data. It is currently a thriving field, as society has become more concentrated on the acquisition, transfer, and exploitation of bioinformatics and biological data.

2. Types of Biological Data

Biological Data can be extracted for use in the domains of omics, bio-imaging, and medical imaging. Life scientists value biological data to provide molecular details in living organisms. Tools for DNA sequencing, gene expression (GE), bio-imaging, neuro-imaging, and brain-machine interfaces are all domains that utilize biological data, and model biological systems with high dimensionality.[1]

Moreover, raw biological sequence data usually refers to DNA, RNA, and amino acids.[1]

Biological Data can also be described as data on biological entities.[2] For instance, characteristics such as: sequences, graphs, geometric information, scalar and vector fields, patterns, constraints, images, and spatial information may all be characterized as biological data, as they describe features of biological beings. In many instances, biological data are associated with several of these categories. For instance, as described in the National Institute of Health's report on *Catalyzing Inquiry at the Interface of Computing and Biology,* a protein structure may be associated with a one-dimensional sequence, a two-dimensional image, and a three dimensional structure, and so on.[2]

CATH - Protein Structure Classification Database. https://handwiki.org/wiki/index.php?curid=1661127

2.1. Biomedical Databases

Biomedical Databases have often been referred to as the databases of Electronic Health Records (EHRs), genomic data in decentralized federal database systems, and biological data, including genomic data, collected from large-scale clinical studies.[3][4]

3. Bio-hacking and Privacy Threats

3.1. Bio-Hacking

Bio-computing attacks have become more common as recent studies have shown that common tools may allow an assailant to synthesize biological information which can be used to hijack information from DNA-analyses.[5] The threat of biohacking has become more apparent as DNA-analysis increases in commonality in fields such as forensic science, clinical research, and genomics.

Biohacking can be carried out by synthesizing malicious DNA and inserted into biological samples. Researchers have established scenarios that demonstrate the threat of biohacking, such as a hacker reaching a biological sample by hiding malicious DNA on common surfaces, such as lab coats, benches, or rubber gloves, which would then contaminate the genetic data.[5]

However, the threat of biohacking may be mitigated by using similar techniques that are used to prevent conventional injection attacks. Clinicians and researchers may mitigate a bio-hack by extracting genetic information from biological samples, and comparing the samples to identify material unknown materials. Studies have shown that comparing genetic information with biological samples, to identify bio-hacking code, has been up to 95% effective in detecting malicious DNA inserts in bio-hacking attacks.[5]

3.2. Genetic Samples as Personal Data

Privacy concerns in genomic research have arises around the notion of whether or not genomic samples contain personal data, or should be regarded as physical matter.[6] Moreover, concerns arise as some countries recognize genomic data as personal data (and apply data protection rules) while other countries regard the samples in terms of physical matter and do not apply the same data protection laws to genomic samples. The forthcoming General Data Protection Regulation (GDPR) has been cited as a potential legal instrument that may better enforce privacy regulations in bio-banking and genomic research.[6]

However, ambiguity surrounding the definition of "personal data" in the text of the GDPR, especially regarding biological data, has led to doubts on whether regulation will be enforced for genetic samples. Article 4(1) states that personal data is defined as "Any information relating to an identified or identifiable natural person ('data subject')"[7]

4. Applications of Deep Learning to Biological Data

As a result of rapid advances in data science and computational power, life scientists have been able to apply data-intensive machine learning methods to biological data, such as deep learning (DL), reinforcement learning (RL), and their combination (deep RL). These methods, alongside increases in data storage and computing, have allowed life scientists to mine biological data and analyze data sets that were previously too large or complex. Deep Learning (DL) and reinforcement learning (RL) have been used in the field of omics research[1] (which includes genomics, proteomics, or metabolomics.) Typically, raw biological sequence data (such as DNA, RNA, and amino acids) is extracted and used to analyze features, functions, structures, and molecular dynamics from the biological data. From that point onwards, different analyses may be performed, such as GE profiling splicing junction prediction, and protein-protein interaction evaluation may all be performed.[1]

Reinforcement learning, a term stemming from behavioral psychology, is a method of problem solving by learning things through trial and error. Reinforcement learning can be applied to biological data, in the field of omics, by using RL to predict bacterial genomes.[8]

Other studies have shown that reinforcement learning can be used to accurately predict biological sequence annotation.[9]

Deep Learning (DL) architectures are also useful in training biological data. For instance, DL architectures that target pixel levels of biological images have been used to identify the process of mitosis in histological images of the breast. DL architectures have also been used to identify nuclei in images of breast cancer cells.[10]

5. Challenges to Data Mining in Biomedical Informatics

5.1. Complexity

The primary problem facing biomedical data models has typically been complexity, as life scientists in clinical settings and biomedical research face the possibility of information overload. However, information overload has often been a debated phenomenon in medical fields.[11] Computational advances have allowed for separate communities to form under different philosophies. For instance, data mining and machine learning researchers search for relevant patterns in biological data, and the architecture does not rely on human intervention. However, there are risks involved when modeling artifacts when human intervention, such as end user comprehension and control, are lessened.[12]

Researchers have pointed out that with increasing health care costs and tremendous amounts of underutilized data, health information technologies may be the key to improving the efficiency and quality of healthcare.[11]

5.2. Database Errors and Abuses

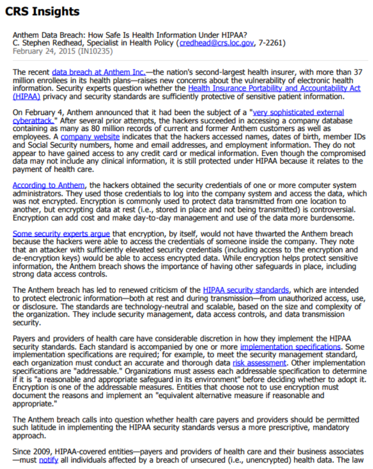
Electronic health records (EHR) can contain genomic data from millions of patients, and the creation of these databases has resulted in both praise and concern.[4]

Legal scholars have pointed towards three primary concerns for increasing litigation pertaining to biomedical databases. First, data contained in biomedical databases may be incorrect or incomplete. Second, systemic biases, which may arise from researcher biases or the nature of the biological data, may threaten the validity of research results. Third, the presence of data mining in biological databases can make it easier for individuals with political, social, or economic agendas to manipulate research findings to sway public opinion.[4][13]

An example of database misuse occurred in 2009 when the *Journal of Psychiatric Research* published a study that associated abortion to psychiatric disorders.[14] The purpose of the study was to analyze associations between abortion history and psychiatric disorders, such as anxiety disorders (including panic disorder, PTSD, and agoraphobia) alongside substance abuse disorders and mood disorders.

However, the study was discredited in 2012 when scientists scrutinized the methodology of the study and found it severely faulty.[15] The researchers had used "national data sets with reproductive history and mental health variables"[14] to produce their findings. However, the researchers had failed to compare women (who had unplanned pregnancies and had abortions) to the group of women who did not have abortions, while focusing on psychiatric problems that occurred after the terminated pregnancies. As a result, the findings which appeared to give scientific credibility, gave rise to several states enacting legislation[16] that required women to seek counseling before abortions, due to the potential of long-term mental health consequences.

Another article, published in the New York Times, demonstrated how Electronic Health Records (EHR) systems could be manipulated by doctors to exaggerate the amount of care they provided for purposes of Medicare reimbursement.[4][17]



CRS Insights

Anthem Data Breach: How Safe Is Health Information Under HIPAA?
C. Stephen Redhead, Specialist in Health Policy (credhead@crs.loc.gov, 7-2261)
February 24, 2015 (IN10235)

The recent data breach at Anthem Inc.—the nation's second-largest health insurer, with more than 37 million enrollees in its health plans—raises new concerns about the vulnerability of electronic health information. Security experts question whether the Health Insurance Portability and Accountability Act (HIPAA) privacy and security standards are sufficiently protective of sensitive patient information.

On February 4, Anthem announced that it had been the subject of a "very sophisticated external cyberattack." After several prior attempts, the hackers succeeded in accessing a company database containing as many as 80 million records of current and former Anthem customers as well as employees. A company website indicates that the hackers accessed names, dates of birth, member IDs and Social Security numbers, home and email addresses, and employment information. They do not appear to have gained access to any credit card or medical information. Even though the compromised data may not include any clinical information, it is still protected under HIPAA because it relates to the payment of health care.

According to Anthem, the hackers obtained the security credentials of one or more computer system administrators. They used those credentials to log into the company system and access the data, which was not encrypted. Encryption is commonly used to protect data transmitted from one location to another, but encrypting data at rest (i.e., stored in place and not being transmitted) is controversial. Encryption can add cost and make day-to-day management and use of the data more burdensome.

Some security experts argue that encryption, by itself, would not have thwarted the Anthem breach because the hackers were able to access the credentials of someone inside the company. They note that an attacker with sufficiently elevated security credentials (including access to the encryption and de-encryption keys) would be able to access encrypted data. While encryption helps protect sensitive information, the Anthem breach shows the importance of having other safeguards in place, including strong data access controls.

The Anthem breach has led to renewed criticism of the HIPAA security standards, which are intended to protect electronic information—both at rest and during transmission—from unauthorized access, use, or disclosure. The standards are technology-neutral and scalable, based on the size and complexity of the organization. They include security management, data access controls, and data transmission security.

Payers and providers of health care have considerable discretion in how they implement the HIPAA security standards. Each standard is accompanied by one or more implementation specifications. Some implementation specifications are required; for example, to meet the security management standard, each organization must conduct an accurate and thorough data risk assessment. Other implementation specifications are "addressable." Organizations must assess each addressable specification to determine if it is "a reasonable and appropriate safeguard in its environment" before deciding whether to adopt it. Encryption is one of the addressable measures. Entities that choose not to use encryption must document the reasons and implement an "equivalent alternative measure if reasonable and appropriate."

The Anthem breach calls into question whether health care payers and providers should be permitted such latitude in implementing the HIPAA security standards versus a more prescriptive, mandatory approach.

Since 2009, HIPAA-covered entities—payers and providers of health care and their business associates—must notify all individuals affected by a breach of unsecured (i.e., unencrypted) health data. The law

A Congressional Research Service report on the safety of health information under HIPAA. Anthem Data Breach: How Safe Is Health Information Under HIPAA? (wikimedia.org)

6. Biomedical Data Sharing

Sharing biomedical data has been touted as an effective way to enhance research reproducibility and scientific discovery.[13][18]

While researchers struggle with technological issues in sharing data, social issues are also a barrier to sharing biological data. For instance, clinicians and researchers face unique challenges to sharing biological or health data within their medical communities, such as privacy concerns and patient privacy laws such as HIPAA.[19]

6.1. Attitudes Towards Data Sharing

According to a 2015 study[19] focusing on the attitudes of practices of clinicians and scientific research staff, a majority of the respondents reported data sharing as important to their work, but signified that their expertise in the subject was low. Of the 190 respondents to the survey, 135 identified themselves as clinical or basic research scientists, and the population of the survey included clinical and basic research scientists in the Intramural Research Program at the National Institute of Health. The study also found that, among the respondents, sharing data directly with other clinicians was a common practice, but the subjects of the study had little practice uploading data to a repository.

Within the field of biomedical research, data sharing has been promoted[20] as an important way for researchers to share and reuse data in order to fully capture the benefits towards personalized and precision medicine.[19]

6.2. Challenges to Data Sharing

Data sharing in healthcare has remained a challenge for several reasons. Despite research advances in data sharing in healthcare, many healthcare organizations remain reluctant or unwilling to release medical data on account of privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA). Moreover, sharing biological data between institutions requires protecting confidentiality for data that may span several organizations. Achieving data syntax and semantic heterogeneity while meeting diverse privacy requirements are all factors that pose barriers to data sharing.[21]

References

1. Mahmud, Mufti; Kaiser, Mohammed Shamim; Hussain, Amir; Vassanelli, Stefano (June 2018). "Applications of Deep Learning and Reinforcement Learning to Biological Data". IEEE Transactions on Neural Networks and Learning Systems 29 (6): 2063–2079. doi:10.1109/tnnls.2018.2790388. ISSN 2162-237X. PMID 29771663. http://dx.doi.org/10.1109/tnnls.2018.2790388.
2. Wooley, John C.; Lin, Herbert S.; Biology, National Research Council (US) Committee on Frontiers at the Interface of Computing and (2005) (in en). On the Nature of Biological Data. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK25464/.
3. Nadkarni, P. M.; Brandt, C.; Frawley, S.; Sayward, F. G.; Einbinder, R.; Zelterman, D.; Schacter, L.; Miller, P. L. (1998-03-01). "Managing Attribute-Value Clinical Trials Data Using the ACT/DB Client-Server Database System" (in en). Journal of the American Medical Informatics Association 5 (2): 139–151. doi:10.1136/jamia.1998.0050139. ISSN 1067-5027. PMID 9524347. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=61285
4. Hoffman, Sharona; Podgurski, Andy (2013). "The use and misuse of biomedical data: is bigger really better?". American Journal of Law & Medicine 39 (4): 497–538. doi:10.1177/009885881303900401. ISSN 0098-8588. PMID 24494442. https://pubmed.ncbi.nlm.nih.gov/24494442/.
5. Islam, Mohd Siblee; Ivanov, S.; Robson, E.; Dooley-Cullinane, T.; Coffey, L.; Doolin, K.; Balasubramaniam, S. (2019). "Genetic similarity of biological samples to counter bio-hacking of DNA-sequencing functionality". Scientific Reports 9 (1): 8684. doi:10.1038/s41598-019-44995-6. PMID 31213619. Bibcode: 2019NatSR...9.8684I. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=6581904
6. Hallinan, Dara; De Hert, Paul (2016), Mittelstadt, Brent Daniel; Floridi, Luciano, eds., "Many Have It Wrong – Samples Do Contain Personal Data: The Data Protection Regulation as a Superior Framework to Protect Donor Interests in Biobanking and Genomic Research" (in en), The Ethics of Biomedical Big Data, Law, Governance and Technology Series (Cham: Springer International Publishing): pp. 119–137, doi:10.1007/978-3-319-33525-4_6, ISBN 978-3-319-33525-4, https://doi.org/10.1007/978-3-319-33525-4_6, retrieved 2020-12-09
7. "Statewatch.org". http://statewatch.org/news/2015/apr/eu-council-dp-reg-4column-2015.pdf.
8. Chuang, Li-Yeh; Tsai, Jui-Hung; Yang, Cheng-Hong (July 2010). "Binary particle swarm optimization for operon prediction". Nucleic Acids Research 38 (12): e128. doi:10.1093/nar/gkq204. ISSN 0305-1048. PMID 20385582. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=2896535
9. Ralha, C. G.; Schneider, H. W.; Walter, M. E. M. T.; Bazzan, A. L. (October 2010). "Reinforcement Learning Method for BioAgents". 2010 Eleventh Brazilian Symposium on Neural Networks: 109–114. doi:10.1109/SBRN.2010.27. ISBN 978-1-4244-8391-4. https://ieeexplore.ieee.org/document/5715222.

10. Xu, Jun; Xiang, Lei; Liu, Qingshan; Gilmore, Hannah; Wu, Jianzhong; Tang, Jinghai; Madabhushi, Anant (January 2016). "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images". IEEE Transactions on Medical Imaging 35 (1): 119–130. doi:10.1109/TMI.2015.2458702. ISSN 0278-0062. PMID 26208307. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=4729702
11. Holzinger, Andreas; Jurisica, Igor (2014), Holzinger, Andreas; Jurisica, Igor, eds., "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions" (in en), Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer): pp. 1–18, doi:10.1007/978-3-662-43968-5_1, ISBN 978-3-662-43968-5, https://doi.org/10.1007/978-3-662-43968-5_1, retrieved 2020-12-09
12. Shneiderman, Ben (March 2002). "Inventing Discovery Tools: Combining Information Visualization with Data Mining". Information Visualization 1 (1): 5–12. doi:10.1057/palgrave.ivs.9500006. ISSN 1473-8716. http://dx.doi.org/10.1057/palgrave.ivs.9500006.
13. Mittelstadt, Brent Daniel; Floridi, Luciano (April 2016). "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts". Science and Engineering Ethics 22 (2): 303–341. doi:10.1007/s11948-015-9652-2. ISSN 1471-5546. PMID 26002496. https://pubmed.ncbi.nlm.nih.gov/26002496/.
14. Coleman, Priscilla K.; Coyle, Catherine T.; Shuping, Martha; Rue, Vincent M. (May 2009). "Induced abortion and anxiety, mood, and substance abuse disorders: isolating the effects of abortion in the national comorbidity survey". Journal of Psychiatric Research 43 (8): 770–776. doi:10.1016/j.jpsychires.2008.10.009. ISSN 1879-1379. PMID 19046750. https://pubmed.ncbi.nlm.nih.gov/19046750/.
15. Kessler, Ronald C.; Schatzberg, Alan F. (March 2012). "Commentary on Abortion Studies of Steinberg and Finer (Social Science & Medicine 2011; 72:72–82) and Coleman (Journal of Psychiatric Research 2009;43:770–6 & Journal of Psychiatric Research 2011;45:1133–4)" (in en). Journal of Psychiatric Research 46 (3): 410–411. doi:10.1016/j.jpsychires.2012.01.021. https://linkinghub.elsevier.com/retrieve/pii/S0022395612000325.
16. "Counseling and Waiting Periods for Abortion" (in en). 2016-03-14. https://www.guttmacher.org/state-policy/explore/counseling-and-waiting-periods-abortion.
17. Abelson, Reed; Creswell, Julie; Palmer, Griff (2012-09-22). "Medicare Bills Rise as Records Turn Electronic (Published 2012)" (in en-US). The New York Times. ISSN 0362-4331. https://www.nytimes.com/2012/09/22/business/medicare-billing-rises-at-hospitals-with-electronic-records.html.
18. Kalkman, Shona; Mostert, Menno; Gerlinger, Christoph; van Delden, Johannes J. M.; van Thiel, Ghislaine J. M. W. (March 28, 2019). "Responsible data sharing in international health research: a systematic review of principles and norms". BMC Medical Ethics 20 (1): 21. doi:10.1186/s12910-019-0359-9. ISSN 1472-6939. PMID 30922290. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=6437875
19. Federer, Lisa M.; Lu, Ya-Ling; Joubert, Douglas J.; Welsh, Judith; Brandys, Barbara (2015-06-24). Kanungo, Jyotshna. ed. "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff" (in en). PLOS ONE 10 (6): e0129506. doi:10.1371/journal.pone.0129506. ISSN 1932-6203. PMID 26107811. Bibcode: 2015PLoSO..1029506F. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=4481309
20. Shneiderman, Ben (2016-07-21). "Inventing Discovery Tools: Combining Information Visualization with Data Mining1" (in en). Information Visualization 1: 5–12. doi:10.1057/palgrave.ivs.9500006. https://journals.sagepub.com/doi/10.1057/palgrave.ivs.9500006.
21. Wimmer, Hayden; Yoon, Victoria Y.; Sugumaran, Vijayan (2016-08-01). "A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy" (in en). Decision Support Systems 88: 51–66. doi:10.1016/j.dss.2016.05.008. ISSN 0167-9236. http://www.sciencedirect.com/science/article/pii/S0167923616300811.

Retrieved from https://encyclopedia.pub/entry/history/show/79965