Application of hybrid algorithms and Explainable Artificial Intelligence ingenomic sequencing

Ángel Canal-Alonso1, Pedro Jiménez1 and Noelia Egido1, Javier Prieto[1], Juan Manuel Corchado[1]

[1]Department of Bioinformatics and Computational Biology, AIR Institute, Carbajosa de la Sagrada, Spain

Email: acanal@air-institute.com

Summary

DNA sequencing is one of the fields that has advanced the most in recent years within clinical genetics and human biology. However, the large amount of data generated through next generation sequencing (NGS) techniques requires advanced data analysis processes that are sometimes complex and beyond the capabilities of clinical staff. Therefore, this work aims to shed light on the possibilities of applying hybrid algorithms and explainable artificial intelligence (XAI) to data obtained through NGS. The suitability of each architecture will be evaluated phase by phase in order to offer final recommendations that allow implementation in clinical sequencing workflows..

Keywords: Next-Generation sequencing, Explainable Artificial Intelligence, Hybrid Algorithms

Introduction

DNA sequencing has revolutionized the field of biology and medicine, providing invaluable tools for deciphering the genetic information underlying all living organisms. As sequencing techniques advance, the volume of data generated has grown exponentially, leading to the need for more advanced and specific techniques for its analysis. It is in this context that artificial intelligence (AI) has begun to play a crucial role, facilitating the processing and interpretation of these vast data sets in reasonable times and with increased accuracy.

Hybrid AI algorithms are those that combine various artificial intelligence techniques or methodologies to achieve more effective results than those that would be obtained by using a single technique. These can combine, for example, deep learning techniques with heuristic or optimization methods, allowing complex and multifaceted problems to be faced, such as genomic sequencing, in a more holistic and robust way.

Hybrid artificial intelligence (AI) algorithms represent a combined approach that integrates different AI techniques and paradigms into a single system, with the goal of overcoming the individual limitations of each technique and ultimately improving overall performance and efficiency. of the resulting algorithm.

In the field of AI, there are numerous methods and techniques, each of which has its strengths and weaknesses. For example, while deep learning algorithms such as neural networks are exceptional at detecting complex patterns in large data sets, they can lack the interpretability offered by more traditional methods such as decision trees. Hybrid algorithms are designed to combine the strengths of multiple techniques, while mitigating their weaknesses.

There are several ways hybrid algorithms can be built:

1. Serial Combination: One technique is used as pre-processing or post-processing for another. For example, you can use a clustering method to segment the data and then apply a classification algorithm on each cluster.

2. Parallel combination: Different techniques work simultaneously and their results are combined. A common example is model assembly, where multiple models (such as decision trees, SVMs, and neural networks) are trained on a single data set, and then their predictions are combined, often by voting, to produce a final result. .
3. Levels of abstraction: One technique takes a high-level view of the problem, while another focuses on more granular details. For example, a neural network can be responsible for extracting general characteristics from the data, while a genetic algorithm could optimize certain parameters based on those characteristics.
4. Knowledge incorporation: Data-driven techniques can be combined with knowledge-based techniques. For example, a system of rules can incorporate expert knowledge in a specific field, while a machine learning algorithm is used to adapt or refine those rules based on real data.

Hybrid algorithms offer great flexibility and are particularly useful in complex problems where there is no single technique that is optimal for all facets of the problem. However, their design and optimization can be more challenging, as interactions between different techniques must be considered and ensure that they work together harmoniously.

On the other hand, explainable AI, also known as XAI (for its acronym in English "Explainable Artificial Intelligence"), refers to methods and techniques in the field of artificial intelligence that produce results that can be easily understood by humans. As AI is increasingly used in critical areas such as medicine and biology, the ability to understand and trust the decisions made by these algorithms is essential. In the context of an NGS pipeline, explainable AI can be crucial to understanding why certain variants have been detected and not others, or to offer insight into possible areas of improvement in the process.

Explainable Artificial Intelligence (XAI) refers to methods and techniques in the domain of artificial intelligence that provide results that are understandable and transparent to humans. As AI models, particularly deep learning models, become more complex and often behave like "black boxes", the need for explainability and transparency has grown in importance.

The main idea behind XAI is that a user has the right to know why a model made a particular decision and how it reached that conclusion. This understanding is essential for several reasons:

1. Trust: For people to trust and widely adopt AI systems, they need to understand how they work and why they make certain decisions. Without a clear explanation, users may be reluctant to trust a system's recommendations or decisions.
2. Accountability and Legal Accountability: In many contexts, such as medical or financial, understanding why a decision was made can be crucial for accountability and regulatory compliance. If something goes wrong, it is essential to identify and explain the cause.
3. Model Improvement: Explainability can help developers identify and correct errors or biases in their models. If you can understand why a model makes errors, it is easier to remedy those problems.
4. Facilitate Adoption: In sectors where decisions have a significant impact on people's lives, such as medicine or transportation, the adoption of new technologies depends largely on their transparency and ability to be audited.

Various techniques and approaches have emerged to address XAI:

1. Intrinsically Explainable Models: These are models that are naturally transparent, such as decision trees or linear regression. Although they are typically less complex and often less accurate than deep learning models, they provide a clear view into how decisions are made.
2. Post hoc methods: These are methods that seek to explain the decisions of a model after it has been trained. A popular example is LIME (Local Interpretable Model-agnostic Explanations), which approximates a complex model with a simpler, locally understandable model around a specific prediction.
3. Visualization: Tools and techniques that allow you to visualize how models work internally, especially neural networks, to give intuition about what characteristics are being considered and how decisions are made.
4. Decomposition Techniques: These techniques decompose predictions into individual feature contributions, providing insight into which features were most relevant to a particular decision.

XAI does not simply seek to create models that are transparent, but rather seeks to ensure that that transparency is useful and meaningful to users, allowing them to make informed decisions and trust AI-based solutions. The growing demand for explainability in AI reflects a broader trend toward ethics and responsibility in technology.

The combination of hybrid algorithms and explainable AI opens an innovative panorama in genomic analysis, allowing not only to increase precision and efficiency in the detection of variants and other related tasks, but also to guarantee that the results are interpretable and justifiable for healthcare professionals. area, thus strengthening confidence in the adoption of these advanced technologies.

Preprocessing phase

Hybrid algorithms combine the features and advantages of multiple approaches to solve a specific problem. In the context of data preprocessing in a Next Generation Sequencing (NGS) pipeline, hybrid algorithms can be extremely valuable in addressing the diversity and complexity of genomic data and improving the efficiency, accuracy, and speed of analysis. Applying the following techniques has been evaluated

1. Error Correction in Sequencing Reads:
    a. Hybrid of k-mer-based techniques and deep learning: By combining traditional k-mer-based methods with recurrent neural networks (RNNs), it is possible to identify and correct errors in sequencing reads more effectively. K-mers can offer a first approximation to detect inconsistencies, while deep learning can refine the correction based on contextual patterns in the sequences.
2. Reading Alignment:
    b. Hybrid of heuristic algorithms and graph techniques: Some algorithms use a combination of heuristic techniques for a first filtering of possible alignments and graph techniques to refine and confirm the alignment. This combination can improve the speed and accuracy of alignment.
3. Elimination of Duplicate Readings:
    c. Hybrid of hashing and machine learning: To identify and remove duplicate reads, a combination of hashing could be used to quickly cluster potentially duplicate reads and machine learning classification algorithms to accurately determine whether the reads are duplicates or simply similar.
4. Normalization of Depth of Coverage:
    d. Hybrid of statistics and generative models: Combining statistical techniques to analyze the distribution of coverage depth together with generative models (such as Generative Adversarial Networks) could allow the generation of synthetic reads that balance coverage depth across regions of the genome.
5. Identification and Trimming of Adapters and Primers:
    e. Hybrid of string search algorithms and deep learning techniques: Identification and removal of adapters and primers is essential to ensure the quality of subsequent analysis. Traditional string search algorithms, such as the Boyer-Moore or Knuth-Morris-Pratt algorithm, can be used initially for rapid identification of known adapter sequences. However, to identify adapters or primers that are not in known databases or that have mutations, a deep learning model could be implemented, specifically a Convolutional Neural Network (CNN), which would be trained with labeled data sets to detect subtle patterns and variations in adapter sequences.
    f. Use of clustering techniques and neural networks: Since adapter sequences may present certain variations, clustering techniques, such as the k-means algorithm, can be applied to group similar reads. Subsequently, with the set of representatives of each cluster, a neural network could be used to refine and confirm the presence of adapters and determine the optimal point for trimming.
    g. Combination of spectral analysis and machine learning-based filtering techniques: Spectral analysis can decompose data sequences into frequency components to detect recurring patterns, such as those introduced by adapters and primers. Once these patterns are detected, machine learning algorithms, such as decision trees or support vector machines (SVM), can be used to classify and filter the reads with adapters, based on features extracted from the spectral analysis.

Incorporating hybrid algorithms in the preprocessing phase of the NGS pipeline is crucial as it can dramatically improve the quality of subsequent genomic analysis. It is essential to combine the strengths of multiple techniques to effectively address the challenges inherent in genomic data and ensure accurate and reliable results.

In the case of explainable artificial intelligence (XAI), it is intended to play a crucial role in ensuring that the decisions made by algorithms are understandable, justifiable and reliable. Here are some examples and approaches to integrate XAI in this phase:

1. Explanation based on feature importance:
    - SHAP (SHapley Additive exPlanations) algorithm: It is a method that is based on game theory to assign a value to each characteristic based on how much it contributes to the decision of a model. For

example, during filtering out low-quality reads, SHAP could reveal which features (such as read length, average quality, presence of adapters, etc.) are most influential in discarding or retaining a read.

2. Deep learning models explained:
   - Grad-CAM (Gradient-weighted Class Activation Mapping): It is useful for convolutional neural networks (CNNs) used in the detection of adapters or contaminants. Grad-CAM can generate heatmaps to highlight areas of the sequence that were vital to the model decision. Thus, biologists and bioinformaticians can get a visual intuition of problem areas in a reading.
3. Intrinsically interpretable models:
   - Decision trees and random forests: These models can be used to classify reads based on quality characteristics and provide clear and understandable decision paths. Although they are not as accurate as deep learning models, their interpretable nature makes them valuable in scenarios where transparency is essential.
4. Explanations based on examples:
   - Counterfactual examples: For a particular read that has been filtered, a counterfactual example can be provided that shows what that read should have been like (in terms of quality, length, etc.) to not be filtered. This gives a clear idea of what factors contributed to its elimination.
5. Explanation through visualization:
   - Tools such as LIME (Local Interpretable Model-agnostic Explanations) can be used to decompose and visualize the decisions of a complex model in simpler and more local terms. For example, if a deep learning-based model decides to discard a read, LIME could identify and visualize the specific regions of the sequence that led to that decision.

Integrating XAI into NGS preprocessing ensures that critical decisions made in this phase are transparent and justifiable. This not only reinforces confidence in the results, but also allows researchers to refine and improve their pipelines based on clear and understandable feedback.

Alignment phase

Sequence alignment is one of the most critical steps in next-generation sequencing (NGS) data analysis. Its main objective is to map the reads obtained from the sequencer to a known genomic reference. Given the enormous amount of data generated in NGS experiments and the inherent complexity of alignment, hybrid algorithms that combine traditional bioinformatics techniques with advanced AI methods can offer significant improvements in accuracy, speed, and efficiency.

1. Combining hashing techniques with neural networks: Traditional alignment algorithms, such as BWA or Bowtie, use hashing techniques to index the reference sequence and facilitate fast searches. We can improve this process by using neural networks to identify error-prone regions or variants in the reads before the alignment process. In this way, the neural network acts as a pre-filter that identifies and corrects errors, which facilitates and accelerates the subsequent hashing and alignment process.
2. Deep learning models for complex region detection: Genomic regions with high repeatability or complex secondary structures often present alignment challenges. A deep learning model, such as a CNN, can be trained to detect these complex regions and reroute the associated reads to more robust or specialized alignment algorithms, while the "simpler" reads can be handled by faster, less intensive algorithms. in resources.
3. Optimizing scoring with machine learning: Traditional algorithms use static scoring matrices, such as BLOSUM or PAM, to guide the alignment process. Using machine learning techniques, we can dynamically adapt these matrices based on specific characteristics of the reads and reference, thereby improving alignment accuracy.
4. Reinforcement learning for alignment decisions: Reinforcement learning can be used to train models that make alignment decisions. In this approach, a reinforcement learning agent is “rewarded” for correct alignment decisions and “penalized” for incorrect decisions, allowing the model to adapt and improve over time.

Applying hybrid algorithms in the alignment phase of an NGS pipeline can provide an optimal combination of speed and accuracy, addressing specific challenges that traditional algorithms may not handle efficiently on their own.

The alignment phase in an NGS pipeline is critical because decisions made during this stage directly affect subsequent phases. The adoption of Artificial Intelligence (AI) has helped improve the accuracy and speed of alignment, but the opaque or "black box" nature of many AI models can be a challenge, especially in genomics, where interpretability and justification of results They are vital. This is where Explainable Artificial Intelligence (XAI) comes in.

1. Visualization of alignment decisions: An XAI-based tool could provide detailed visualizations that explain how and why certain reads aligned to specific locations. For example, highlight features in the read sequence that had significant weight in the alignment decision. This is essential so that bioinformaticians can validate the accuracy of the alignment and detect possible errors or anomalies.
2. Interpreting scoring matrices: Although scoring matrices, such as BLOSUM, are static, when incorporating machine learning techniques that adapt these matrices, XAI can provide insights into which characteristics of readings and references are most influencing adaptation of these matrices.
3. Decomposing complex genomic regions: When a deep learning model identifies difficult or complex genomic regions for alignment, XAI can decompose and show what specific features make these regions challenging. This could include visualizations that highlight repetitiveness, secondary structures, or mutations.
4. Feedback and continuous learning: By offering clear and understandable explanations of alignment decisions, XAI allows for more effective feedback from genomics experts. They can quickly identify where the AI model might be making mistakes and provide corrections. This feedback is crucial for the continuous training and refinement of the model.
5. Building trust: Genomics is a discipline where errors can have significant implications, especially in clinical applications. By providing transparency into the alignment process, XAI builds trust among researchers, clinicians, and other stakeholders, assuring them that AI-based alignment decisions are robust and reliable.

In summary, while AI has revolutionized the alignment phase in NGS with its ability to quickly process large volumes of data and make accurate decisions, XAI complements these advances by providing the transparency and understanding necessary to continually validate, correct and improve these processes. of alignment.

Variant detection phase

Variant detection is one of the most critical and challenging aspects in next-generation sequencing (NGS) analysis. It is essential to identify variations in DNA, such as SNPs (single nucleotide polymorphisms), indels (insertions or deletions), and structural variants with high precision and sensitivity. The application of hybrid algorithms in this phase combines traditional rule-based techniques with machine learning to offer a more robust and adaptive approach.

1. Incorporation of prior knowledge: Traditional algorithms for variant detection are based on prior knowledge and rules defined by genomics experts. These hybrid algorithms can combine this knowledge with machine learning models that are trained on large data sets. For example, a rule-based model could be used to identify variant candidates, and then a deep learning model to refine and validate these candidates.
2. Variant classification models: Hybrid algorithms can combine heuristic features with features learned by machine learning models. For example, a combination of statistical features, such as base quality, depth of coverage, and allele balance, along with features learned by neural networks, can be used to classify variants more accurately.
3. Predicting functional effects: Once variants are identified, it is crucial to predict their functional impact. Hybrid algorithms can combine annotated databases with machine learning techniques to predict the effect of a variant on gene or protein function.
4. Integration of multiple data sources: DNA sequencing is rarely performed in a vacuum. Often, other data is available, such as gene expression data, epigenetic information, or phenotypic data. Hybrid algorithms are particularly suitable for integrating these multiple data sources, using traditional techniques to process high-quality data and machine learning for heterogeneous or incomplete data.
5. Adaptive models: One of the key advantages of hybrid algorithms is their ability to adapt to new data. As more genomes are sequenced and new variants are discovered, models can be retrained and adapted, combining prior knowledge with new insights.

With the increasing adoption of machine learning techniques in variant detection, there is a need to understand and explain the decisions made by these complex models. This is where Explainable Artificial Intelligence (XAI) comes into play.

1. Explaining complex model decisions: When using machine learning algorithms, such as neural networks, to identify genomic variants, it is vital to understand how and why a model made a particular decision. XAI can help visualize and break down the contributions of specific features (such as read quality, depth of coverage, etc.) to the final model decision. This allows geneticists and clinicians to have confidence in the results and also identify areas where the model could be improved.
2. Local and global interpretation: XAI techniques can not only offer insights into individual decisions (local interpretation) but also into the overall behavior of

the model (global interpretation). For example, you can understand which features are generally the most informative for variant detection or whether certain genomic regions tend to be more problematic for the model.

3. Model validation and refinement: The explanations provided by XAI can be used to validate and refine models. If a model is making decisions based on features that have no biological relevance, this may be an indication that it needs to be retrained or tuned.
4. Improved interdisciplinary communication and collaboration: XAI not only benefits those who develop the models but also a broader audience of users. The generated visualizations and explanations facilitate communication between bioinformaticians, geneticists, clinicians and other healthcare professionals, allowing for more informed discussions about genomic findings and their clinical relevance.
5. Facilitate clinical decisions: In a clinical context, decisions based on variant detection can have significant implications for the diagnosis, prognosis, and treatment of diseases. XAI's ability to explain why a variant was classified as pathogenic or benign can help clinicians make more informed decisions.

Variant annotation phase

Annotating variants in an NGS pipeline involves associating biological, functional and clinical information with the identified variants. The goal is to understand the potential biological impact of a variant and its clinical relevance, if any. Hybrid algorithms, which combine symbolic and sub-symbolic learning techniques, can play a crucial role in this phase to optimize the accuracy and robustness of the annotation.

1. Integration of Different Data Sources: Variants can be annotated using multiple databases and biological resources. Hybrid algorithms allow the integration of this diversity of sources, by using the capacity of sub-symbolic learning to handle unstructured or semi-structured data and symbolic logic to integrate domain-specific rules and knowledge.
2. Variant Prioritization: Not all variants identified in a genome are clinically relevant. Hybrid algorithms can employ deep learning techniques to recognize patterns in the data and symbolic techniques to apply biological and clinical rules, resulting in more accurate prioritization of variants based on their relevance and potential pathogenicity.
3. Optimizing Functional Interpretation: The function of a genetic variant can be inferred from multiple characteristics, such as its location, evolutionary conservation and effect on the protein. Here, the symbolic component can provide a logical structure for combining these features, while the sub-symbolic component can capture complex and subtle interactions between them.

This project proposes to investigate the following architectures in greater depth:

1. Hybrid of Neural Networks and Rule-Based Systems: An architecture could employ a neural network to extract features from genomic sequences and patterns in annotation databases. From there, a rules-based system could step in to assign specific labels to variants based on their clinical relevance, using standard criteria and guidelines.
2. Model based on Decision Trees and Convolutional Networks: A decision tree can segment the annotation process into different stages depending on the type of variant and its location. At each stage, a convolutional network could analyze the local sequence around the variant to identify specific functional features.
3. Integrating Fuzzy Logic with Deep Networks: In situations where there is ambiguity in the annotation, fuzzy logic can help capture this uncertainty. For example, if a variant is possibly but not definitely pathogenic, a fuzzy logic algorithm could provide a degree of certainty. A deep network, on the other hand, could be used to determine this degree based on multiple characteristics of the variant and its context.

These examples demonstrate the potential of hybrid algorithms in the variant annotation phase. By combining different techniques, these algorithms can offer robust and accurate solutions to interpret the biological and clinical meaning of genetic variants.

Explainable Artificial Intelligence (XAI) aims to design AI models that offer clear and understandable explanations of their decisions, which is essential to ensure trust in the annotations provided by AI algorithms.

1. Rationale for Annotations: It is vital that researchers and clinicians understand why a variant has been annotated in a certain way. XAI can provide clear reasoning, such as identifying specific features of a variant that indicate its potential pathogenicity or highlighting prior knowledge that supports a particular annotation.
2. Validating Annotations with Domain-Based Rules: Explainable models can be designed to provide justifications based on rules established in the scientific literature, ensuring that annotations are aligned with current knowledge in genomics.

3. Visualization of Important Features: XAI can provide visualizations that highlight genomic regions or features that had a significant impact on the annotation decision. This is especially useful for identifying conserved regions, functional domains, or binding sites that may influence the function of a variant.

We propose as models to apply within the project:

1. Attention Models in Neural Networks: These models can highlight specific areas of a genomic sequence that were key to an annotation decision. For example, if a variant is annotated as pathogenic due to its location in a conserved domain, the attention model can visualize and highlight that region as justification.
2. Interpretable Decision Trees: Decision trees are inherently explainable and can be designed to make annotation decisions based on clear and justifiable criteria, such as the presence of a variant in a specific region or its comparison with databases of known variants.
3. Attribution Decomposition Methods: These methods, such as SHAP (SHapley Additive exPlanations), can decompose the contribution of different characteristics to the final decision of the model. In the context of annotation, this could translate into providing weights or importances to different genomic regions or databases queried during the annotation process.
4. Interactive Interfaces: XAI-based tools can provide interfaces where users can interact with the model, adjust parameters, and observe how these changes affect annotation decisions, giving them a deeper understanding of the process.

By implementing XAI in the variant annotation phase, it can be ensured that decisions made by complex algorithms are transparent, defensible, and aligned with expert knowledge in the genomic field. This transparency is essential to gaining the trust of researchers and medical professionals who depend on accurate annotations to make informed decisions.

Conclusions

As we move into the era of precision genomics, the need for sophisticated tools and algorithms that can handle, analyze and deliver insights from large genomic data sets has become evident. In this context, hybrid AI algorithms and Explainable Artificial Intelligence (XAI) have proven to be essential tools in the processing, alignment, search and annotation of variants within an NGS pipeline.

Hybrid algorithms combine the best of two worlds: the predictive capabilities of deep learning models and the structure and understandability of traditional algorithms. This combination has significantly improved accuracy and efficiency in various phases of the pipeline, from preprocessing to variant annotation. However, with the increasing complexity of these models, transparency and interpretability have become central concerns, and this is where XAI comes in.

Incorporating XAI into genomic analysis not only strengthens the accuracy of the models, but also provides an additional layer of transparency, ensuring that decisions made by the algorithms can be understood, justified and, ultimately, trusted. This transparency is vital to ensure the acceptance and trust of medical professionals, researchers and patients in the results provided by an AI-based pipeline.

In summary, as we embark on deeper genomic discoveries and seek to translate these findings into effective clinical interventions, the integration of hybrid algorithms and XAI into NGS pipelines will be essential. These tools will not only boost the accuracy and efficiency of genomic analysis, but will also ensure that the insights generated are transparent, interpretable and, most importantly, reliable. The combination of computational power and explainability puts us in an optimal position to take full advantage of the benefits of the genomic revolution in the coming years.

Acknowlegments

This study has been funded by the AIR Genomics project (with file number CCTT3/20/SA/0003), through the call 2020 R&D PROJECTS ORIENTED TO THE EXCELLENCE AND COMPETITIVE IMPROVEMENT OF THE CCTT by the Institute of Business Competitiveness of Castilla y León and FEDER funds

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-

Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6