Deep Symbolic Learning Architecture for Variant Calling in NGS

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

The Variant Detection process (Variant Calling) is fundamental in bioinformatics, demanding maximum precision and reliability. This study examines an innovative integration strategy between a traditional pipeline developed in-house and an advanced Intelligent System (IS). Although the original pipeline already had tools based on traditional algorithms, it had limitations, particularly in the detection of rare or unknown variants. Therefore, SI was introduced with the aim of providing an additional layer of analysis, capitalizing on deep and symbolic learning techniques to improve and enhance previous detections.

The main technical challenge lay in interoperability. To overcome this, NextFlow, a scripting language designed to manage complex bioinformatics workflows, was employed. Through NextFlow, communication and efficient data transfer between the original pipeline and the SI were facilitated, thus guaranteeing compatibility and reproducibility.

After the Variant Calling process of the original system, the results were transmitted to the SI, where a meticulous sequence of analysis was implemented, from preprocessing to data fusion. As a result, an optimized set of variants was generated that was integrated with previous results. Variants corroborated by both tools were considered to be of high reliability, while discrepancies indicated areas for detailed investigations.

The product of this integration advanced to subsequent stages of the pipeline, usually annotation or interpretation, contextualizing the variants from biological and clinical perspectives. This adaptation not only maintained the original functionalities of the pipeline, but was also enhanced with the SI, establishing a new standard in the Variant Calling process. This research offers a robust and efficient model for the detection and analysis of genomic variants, highlighting the promise and applicability of blended learning in bioinformatics.

Keywords: Next-Generation sequencing, Explainable Artificial Intelligence, Deep Symbolic Learning

---

Introduction

Deep Symbolic Learning (ASP) represents an innovative trend in the field of artificial intelligence, focusing on fusing the advantages of Deep Learning with those of symbolic learning. This method was born in response to one of the main limitations of Deep Learning: the absence of clarity and explainability in its models. What ASP seeks is to create systems that, in addition to being highly efficient in their performance, are equally clear and justified in their operations.

Deep Symbolic Learning works by combining neural networks, which are adept at learning detailed and stratified representations of information, with symbolic schemes that favor the development of logical models and well-defined semantics. Unlike the purely numerical or statistical approach

of traditional Deep Learning, ASP integrates elements such as symbols, guidelines and logical links in its process, thus giving an additional dimension of context and organization to the knowledge generated.

One of the outstanding points of ASP is its ability to capitalize on knowledge previously encoded in symbolic formats, using it as a basis to guide and enrich the learning of neural networks. This characteristic is essential in areas where there is a vast body of knowledge, as is the case in fields such as biology and genomics.

Within the Variant Detection phase in Next Generation Sequencing (NGS) workflows, ASP emerges as a strategy with a lot of potential. Neural networks have the ability to discern intricate patterns and nuances in sequencing data. In parallel, the symbolic component can include guidelines and previously known data on genetic variants, mutations, and their biological importance. This could not only increase the accuracy of variant detection, but also provide logical and prior knowledge-based rationales for the reasons for classifying certain sequences as variants.

The duality inherent in Deep Symbolic Learning also introduces superior versatility in modeling. While neural networks adapt to specificities and variations in information, the symbolic segment can function as a bulwark, ensuring that inferences and results are in tune with the previously established biological knowledge base. In this way, the ASP is consolidated as a cutting-edge and reliable tool to address the specific challenges of variant detection and the genomic field as a whole.

In the field of bioinformatics and genomics, precision, efficiency and explainability are essential imperatives. Responding to these needs, this study presents the conception and structuring of an Intelligent System based on Deep Symbolic Learning (ASP) with the primary objective of optimizing the Variant Detection phase in Next Generation Sequencing (NGS) workflows. This innovation is the fruit of an extensive and meticulous design and experimentation process.

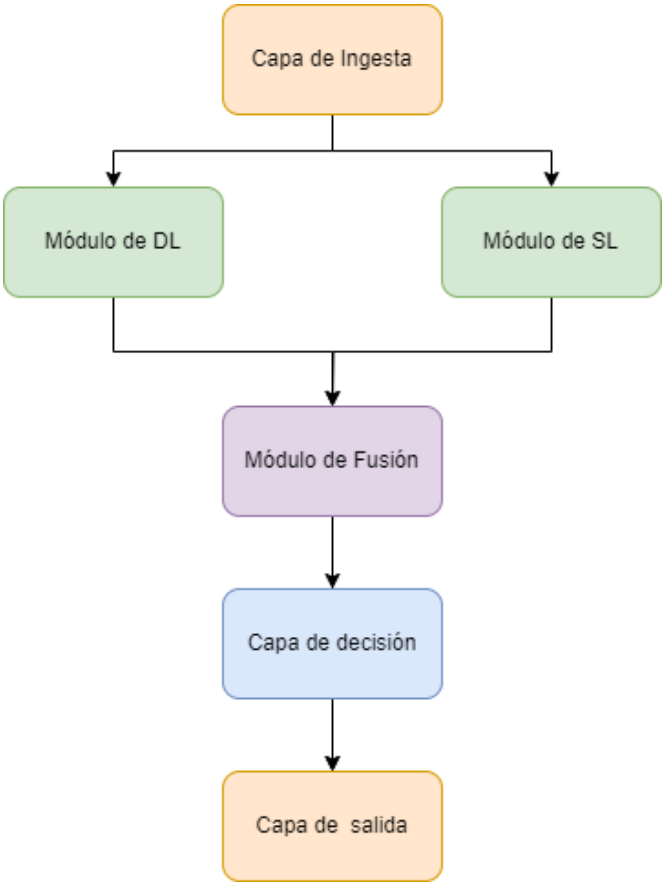
The underlying motivation for developing such a system lies in the convergence of two paradigms: the power and adaptability of deep learning and the semantic and structural clarity of symbolic learning. By amalgamating these two perspectives, we aspire to configure a tool that not only boasts cutting-edge precision but also provides scientists and health professionals with a transparent and interpretable perspective of the derived results.

The system architecture is deployed through different crucial stages. Initially, it focuses on the acquisition and rigorous preprocessing of the sequencing data, guaranteeing the quality and relevance of the information that will be fed to the model. Subsequently, we embarked on the design, construction and training of hybrid models, which take advantage of the depth of neural networks and the semantics of symbolic representations. Finally, it focuses on the interpretation of the results, not only from a numerical or statistical perspective, but also from a logical and semantic framework.

Each of these segments has been finely calibrated to ensure maximum cohesion between the deep and symbolic learning components. This integration ensures that, regardless of the variations and noises intrinsic to the sequencing data, the system not only retains its robustness but also remains aligned with the corpus of consolidated genomic knowledge. Taken together, this work aims to set a new standard in the field of genomic variant detection, advocating an integrative approach that combines precision with explainability.

Model architecture

The model architecture represents the core of the system and has been designed to efficiently encapsulate the capabilities of Deep Symbolic Learning. This architecture is made up of multiple layers and modules, each with a specific function, that work together to interpret sequencing data.



1. Ingestion and Preprocessing Layer: Before any model can work with the data, it must be properly prepared. This layer handles raw NGS data ingestion

and performs essential preprocessing operations such as normalization, error correction, and preliminary identification of regions of interest. Furthermore, it transforms the data into representations that can be easily consumed by subsequent layers.

2. Deep Learning Module: Once preprocessed, the data is fed to this module, which consists of convolutional and recurrent neural networks designed specifically for genomic sequences. These networks detect patterns and features in the data, such as conserved sequences, repetitive regions, or potential variant sites. The deep nature of this module allows the system to learn hierarchical representations of the data, from low-level features such as nucleotide trios to high-level abstractions such as genomic structures.
3. Symbolic Learning Module: Parallel to the deep learning module, the symbolic module operates on the same data. Using a predefined set of rules, heuristics, and prior knowledge of genomics, this module identifies and validates findings. For example, you can confirm the identification of a variant if it matches known rules about mutation patterns or specific genomic contexts.
4. Fusion Module: This is one of the most critical components of the architecture. Here, the outputs of the deep and symbolic learning modules are combined and integrated into a unified representation. Data fusion techniques are used to combine the strengths of both approaches, ensuring that the system not only detects patterns in the data but also validates and interprets them according to existing genomic knowledge.
5. Decision Layer: With the unified information from the fusion module, this layer is responsible for making final decisions about the identification and classification of variants. It uses both quantitative information (such as the probability assigned by the deep learning module) and qualitative information (such as the validation provided by the symbolic module) to make final judgments.
6. Interpretation and Output Layer: Once the decisions are made, this layer presents the results in a way that can be interpreted by the user. Detected variants are annotated with relevant information, such as their potential functional or clinical impact. Additionally, thanks to the symbolic component, the system can provide clear explanations as to why a particular variant was identified.

As a whole, the architecture is designed to be modular, allowing individual components to be updated or replaced as research progresses or needs change. This flexibility ensures that the system can keep up with rapid advances in genomics and bioinformatics.

Data preprocessing

The preprocessing process in the context of next generation sequencing (NGS) is crucial, as it largely determines the quality and accuracy of subsequent results. Despite advances in sequencing technologies, data obtained directly from instruments are often plagued by imperfections, artifacts, and noise. The Ingestion and Preprocessing Layer acts as the first filter and transformer of this raw data, preparing it for subsequent analysis.

Initially, when the data is ingested, the layer begins by identifying and separating the different samples and runs, ensuring that each set of sequences is clearly demarcated and catalogued. This initial organization is essential to avoid confusing or mixing data from different sources or experimental conditions. Once cataloged, the data goes through a quality filtering process. Here, sequences that do not meet a specific quality threshold, determined by read quality, are discarded or corrected. This correction is done using redundant information in the data, such as overlapping reads, and with the help of specialized techniques that estimate the correct sequence based on patterns observed in the data.

Next, an alignment step is carried out, where the sequences are mapped against a reference genome. It is in this phase where variations, such as SNPs and indels, begin to stand out. However, due to the inherently noisy nature of sequencing data, not all observed discrepancies from the reference genome are due to real variants. Systematic errors in sequencing, artifacts and noise can create false signals. Therefore, a post-alignment refinement step that recalibrates and optimizes the quality of the alignment, adjusting and correcting possible errors, is essential.

Once the data has been aligned and refined, an operation to identify regions of interest is performed. These regions are segments of the genome that are particularly relevant for subsequent analysis, either because they show signs of variation, because they are associated with known genomic regions of clinical or functional interest, or because they exhibit patterns that indicate the presence of regulatory elements or other characteristics. genomics.

In parallel with the identification of these regions, the layer is also dedicated to normalizing the data. Normalization is essential to ensure that data from different sources, technologies or experimental conditions are comparable to each other. This task may involve adjusting the depth of coverage, correcting systematic biases, or transforming the sequences into representations that are friendlier for subsequent analysis.

The result of this preprocessing layer is a clean, structured data set ready to be consumed by the deep and symbolic learning modules. It is essential to understand that, despite being only the first step in the pipeline, the decisions and operations carried out in this phase have a direct impact on the

precision, efficiency and quality of the final results of the system.

Model parameters

Deep Learning Module

The Deep Learning Module, as an integral part of our system, plays an essential role in detecting and understanding complex and subtle patterns in sequencing data. This detection relies on deep learning's ability to extract hierarchical features from data, from fundamental aspects such as individual nucleotide sequences to higher-level interpretations such as the identification of genomic structures and regulatory regions.

One of the cornerstones of the module is the incorporation of convolutional neural networks (CNNs). CNNs are particularly suitable for processing sequencing data, as they can recognize local patterns within sequences and are shift invariant. That is, if a particular nucleotide sequence is indicative of a variant or a genomic phenomenon, CNNs can detect it regardless of its position in the input sequence. Convolutional layers in the network examine overlapping segments of the genomic sequence, identifying relevant patterns and features through filters, and transforming them into high-level representations that are easier to interpret by subsequent layers.

On the other hand, recurrent neural networks (RNNs) are also integrated, and more specifically LSTM (Long Short-Term Memory) given the sequential nature of the genomic data. These networks are adept at handling long-term dependencies and variable length sequences, making them ideal for understanding broader contexts in sequencing data. For example, the influence of a distant sequence on the expression or function of a particular gene can be captured by these recurrent units.

While CNNs focus on spatial patterns and RNNs capture temporal dynamics, the combination of both allows the module to obtain a complete and detailed view of the data. This synergy between convolutional and recurrent networks is crucial for addressing the complexities and variabilities inherent in genomic data.

Additionally, in the training process of this module, it is essential to define and adjust various parameters. The learning rate, batch size, loss function and regularization are critical aspects that determine the effectiveness of the model in learning from the data without falling into overfitting. The model is trained using large labeled data sets, where genomic sequences are accompanied by information on known variants, genomic structures and other relevant annotations. Over time, the model adjusts its weights and internal parameters to minimize the discrepancy between its predictions and the actual data, arriving at an optimal representation that can generalize to new data not previously seen.

Taken together, the Deep Learning Module is an amalgamation of advanced machine learning techniques that, when working together, allow the system to identify, classify and understand the variants and structures present in next generation sequencing data with precision and efficiency. without precedents. The adaptive and evolutionary nature of this module ensures that, as more data and knowledge becomes available, the system can continue to improve and refine its genomic interpretation capabilities.

Symbolic learning module

The Symbolic Learning Module represents a fundamental facet of our system that contrasts and complements the deep learning module. While deep learning digs into data to discover implicit patterns and complex relationships, symbolic learning focuses on representing and using explicit, structured knowledge about the domain in question, in this case, genomics.

The basis of this module lies in the creation and manipulation of symbolic representations of information. In the context of genomic sequencing, these representations address known genomic structures, rules inherited from previous studies, genetic patterns associated with specific phenotypes, among others. These symbols and rules are organized into knowledge structures, often referred to as knowledge bases, which are essentially systems of rules or logic designed to reason about data.

One of the central approaches within symbolic learning is the rule-based system. These rules are derived from prior knowledge, scientific literature or even through experts in the field. These rules are applied to the data to filter, classify and predict the presence of particular genomic phenomena. The rules defined for this model are:

- Motif Sequence Rule: If a specific nucleotide motif is detected in a promoter region, it predicts the binding of a known transcription factor.
- Pathogenic Variant Rule: If a variant is identified in an exon of a gene associated with an inherited disease and that variant has previously been classified as pathogenic, classify it as high risk.
- Splicing Rule: If a variant is detected in the first or last two positions of an intron, consider the possibility that it affects the splicing sites and, therefore, the formation of the mRNA.
- Conservation Rule: If a variant is found in a highly conserved region across different species, that region likely has an important biological function.
- Repetition Rule: If a sequence has multiple repeats of a specific trinucleotide, consider the possibility that it is related to trinucleotide repeat diseases.

- Silencer Rule: If a variant is found in a region known to contain silencing elements, evaluate the potential of said variant to affect gene regulation.
- Protein Interaction Rule: If a variant is identified in a protein interaction domain, investigate its potential impact on the formation of protein complexes.
- Synonym Rule: If a variant does not change the resulting amino acid in a protein, it is generally classified as synonymous, but can still be reviewed for potential effects on splicing or mRNA stability.
- Founder Effect Rule: If a specific variant is common in a particular population or ethnic group and is associated with a disease, consider the possibility of a founder effect.
- Compensation Rule: If multiple variants are detected in the same gene or pathway and one is pathogenic, investigate the other variants to see if they have a potential compensatory effect.

A key advantage of symbolic learning is its ability to interpret. Unlike deep learning models, which are often considered black boxes, rule-based systems offer clear and transparent reasoning behind each decision or prediction. This clarity is invaluable in fields such as genomics, where the interpretation and justification of results can have significant implications in areas such as clinical diagnosis or biomedical research.

However, it is not enough to simply codify existing knowledge. The module is also capable of "learning" or refining its rules and representations based on new data. Using techniques such as rule induction, the module can examine the data, compare it to its current knowledge base, and adjust, delete, or create new rules to better reflect the reality of the data. This is especially useful in an ever-evolving field like genomics, where new discoveries can change our understanding of biological systems.

Combined with the Deep Learning Module, the Symbolic Learning Module offers a holistic and deep understanding of genomic data. While the former focuses on discovering non-obvious patterns and relationships in the data, the latter provides a structured and justified framework for interpreting and reasoning about these discoveries. Together, they offer a powerful combination of data-driven intuition and knowledge-based reasoning, enabling the system to operate with accuracy, efficiency and transparency unmatched in next-generation sequencing analysis.

Fusion Module

The Fusion Module stands as the essential integrative component in our system, responsible for amalgamating the results and intuitions obtained from both the Deep Learning Module and the Symbolic Learning Module. This task is essential since, although both modules separately are powerful, it is their coordinated collaboration that gives rise to the true synergy and enhancement of the analysis.

From a technical perspective, the Fusion Module operates in several stages. Initially, it collects the outputs of the Deep Learning Module. These outputs, in the form of feature vectors, latent representations, or direct classifications, encapsulate complex patterns and nonlinear relationships discovered in the data. These representations are extremely valuable but may lack direct interpretability or connections to explicit biological knowledge.

Simultaneously, the Fusion Module accesses the knowledge base of the Symbolic Learning Module. Rules, structures, and symbolic representations provide a structured framework and contextualization for data, based on decades of research and understanding in genomics.

With both sources of information at its disposal, the Fusion Module begins the integration process. It uses advanced techniques, such as reasoning based on fuzzy logic and attention neural networks, to appropriately weight the information from both modules. In essence, it is about determining where to trust data-driven predictions most and where to apply symbolic knowledge to correct, guide or complement those predictions.

For example, if the Deep Learning Module detects a possible genetic variant of interest but that variant contradicts a well-established symbolic rule, the Fusion Module can choose to prioritize the rule or at least send an alert to a more detailed review.

A crucial consideration in this process is feedback. The Fusion Module not only integrates, but also learns. As it receives more data and faces more scenarios, it refines its ability to balance and combine information from the other modules. This is essential to ensure that the system, as a whole, remains adaptive and evolutionary, adjusting to new challenges and discoveries in the field of genomics.

Finally, the Fusion Module completes its operation by producing a series of unified results that incorporate both deep and symbolic learning. These results may be classifications, predictions, annotations or any other output format relevant to genomic analysis, but what is certain is that they reflect an integrated and holistic view of the problem, taking advantage of the best of both worlds.

Final implementation

In the constantly evolving environment of bioinformatics, the Variant Detection process (Variant Calling) represents a critical pillar, where precision and reliability are essential. In this context, the original pipeline, developed in-house, already had advanced tools for this process. Although these tools, based on conventional algorithms and reference repositories, demonstrated reasonable efficiency, they presented certain limitations, particularly in terms of false positives and in the detection of atypical or not yet cataloged variants.

In order to reinforce and not supplant the existing system, an Intelligent System (SI) was introduced. This strategy was motivated by the aspiration that the SI would offer an additional level of analysis, using deep and symbolic learning techniques, in order to review and potentially improve the detections previously identified by the original pipeline.

The first technological challenge faced was to guarantee fluid and effective communication between the SI and the tools of the original pipeline, given the particularities of both systems. In this scenario, we turned to NextFlow, a scripting language designed specifically to manage complex bioinformatics workflows. Thanks to its innate capabilities to coordinate, monitor and ensure reproducibility of tasks, NextFlow emerged as the ideal platform for our integration.

The integration process began with the creation of a specific function in NextFlow that invoked the SI. Subsequently, after the Variant Calling process of the original system, the transfer of results to the SI was facilitated using NextFlow channels. This transition was managed by ensuring that the data was compatible, generally adopting the VCF format.

Once inside the SI, an analysis sequence was established that began with a preprocessing module, followed by deep and symbolic learning modules, culminating with a fusion module. The culmination of this process generated a set of optimized variants that were combined with previous results.

This convergence offered notable benefits. Variants identified uniformly by both tools were considered high reliability. The discrepancies, by contrast, provided an indication for future research or detailed reviews.

The final product of this integration was transmitted to the next stage of the pipeline, usually oriented towards annotation or interpretation, where the variants were analyzed from a biological and clinical perspective.

In summary, this adaptation allowed the original pipeline to not only retain its initial functionalities, but also to be enriched with the depth and precision of the SI, thus enhancing the comprehensive Variant Calling process. The integration described in this study can serve as a paradigm for future research and development in the bioinformatics field.

Conclusions

Despite the success and notable improvement that the Intelligent System (IS) has brought to the Variant Calling process, as with any emerging technology, there are certain limitations that must be addressed to reach its full potential.

One of the main restrictions has been computational capacity. Although SI is highly efficient in its operation, deep and symbolic learning modules, by their nature, require a high degree of computational power, especially when handling large volumes of data. At the current stage, it has occasionally faced bottlenecks in terms of processing speed. With additional investment in infrastructure, such as purchasing more powerful hardware or deploying to cloud computing environments with greater resources, these challenges could be easily mitigated.

Another limitation has been the size and diversity of the training data available. While vast data sets have been used to train the SI, there is always the risk of inadvertent bias. Expanding the diversity of the data, both in geographic and ethnic terms, would improve the generalization of the system to different populations. Additional funding could go toward acquiring more data and establishing collaborations with institutions that hold diverse genomic repositories.

Integration with external databases, although it has been a strength, also has limitations. Depending on the availability and updating of these databases, there could be gaps in the knowledge that the IS can access. In the future, it would be ideal to consider the creation of an internal database, constantly updated, that compiles the most recent information in genomics and that can be fed by both internal and external results.

Finally, while incremental retraining has been an advantage, the periodicity and effectiveness of retraining could improve with the implementation of a more automated system that continually monitors the emergence of new data and adjusts the model in real time. This could require, again, an investment in development and cutting-edge monitoring systems.

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

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