DESENVOLVIMENTO DE UMA BASE DE ELEMENTOS TRANSPONÍVEIS DO GENOMA PARA APLICAÇÃO EM TREINAMENTO E TESTES DE REDES NEURAIS ARTIFICIAIS

Caroline Heloíse de Oliveira (BIC/FA), Gabriel Carneiro de Arruda, Adriano Ferrasa (Orientador), e-mail: ferrasa@uepg.br.

Universidade Estadual de Ponta Grossa, Departamento de Informática.

Ciência da Computação (1.03.00.00-7) / Metodologia e Técnicas da Computação (1.03.03.00-6)

Palavras-Chave: *elementos transponíveis, dataset, base de treinamento, redes neurais artificiais, inteligência artificial.*

Resumo

Este projeto aborda a classificação de elementos transponíveis (TEs) através do desenvolvimento de um *dataset* otimizado para Redes Neurais Artificiais (RNAs). A partir de bases públicas, sequências foram extraídas, categorizadas por superfamília e convertidas para um formato binário. O *dataset* final, com 335.595 sequências, foi validado com uma Rede Neural Convolucional (CNN), confirmando sua funcionalidade. O trabalho entrega um recurso robusto para treinar modelos de aprendizagem profunda, visando aprimorar a anotação genômica automatizada de TEs.

Introdução

TEs, descritos por McClintock (1950), são sequências de DNA móveis no genoma. Antes considerados "DNA lixo" (Feschotte, 2008; Mita; Boeke, 2016), hoje são reconhecidos como fontes de variabilidade genética e associados a doenças. A complexidade em classificá-los impulsionou o uso de RNAs, algoritmos de aprendizagem profunda que reconhecem padrões em dados genômicos (Heaton, 2018). Este projeto visou criar uma base de dados de TEs, robusta e de alta qualidade, para otimizar o treinamento e teste de RNAs e, assim, aprimorar a classificação destes importantes elementos genômicos.

Materiais e Métodos

A partir de uma revisão sistemática da literatura, foram selecionadas as bases de dados públicas DPTEdb, RepBase, RiTE e TEfam. Uma rotina computacional foi desenvolvida para automatizar a extração e organização das sequências em arquivos FASTA distintos por superfamília. Em seguida, os dados foram convertidos para um formato numérico através de uma representação binária (mapeamento de 2 bits). Por fim, a funcionalidade e coesão da base de dados foram validadas pela implementação de uma CNN.

Resultados e Discussão

O processamento resultou em um *dataset* final com 335.595 sequências de TEs, categorizadas em 29 superfamílias. Este volume de dados oferece uma base estatisticamente robusta, essencial para treinar modelos de aprendizagem profunda. A organização por superfamília é um diferencial, permitindo o desenvolvimento de classificadores especializados. A validação com uma CNN confirmou que o formato binário é eficiente e que as sequências contêm padrões aprendíveis, atestando a funcionalidade e o valor prático da base gerada.

Conclusão ou Conclusões

O projeto cumpriu seu objetivo ao gerar uma base de dados robusta e funcional para o treinamento de RNAs. A contribuição principal é a entrega de um recurso prático que aborda a carência de dados de treinamento bem estruturados, um gargalo para o avanço da área. Foi validada uma metodologia replicável para a criação de *datasets* genômicos customizados. A base de dados gerada tem o potencial de catalisar novas pesquisas, acelerando o desenvolvimento de ferramentas mais acuradas para a anotação genômica de Tes.

Agradecimentos

Os autores agradecem à Fundação Araucária pelo apoio financeiro.

Referências

FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. **Nature Reviews Genetics**, [s. l.], v. 9, p. 397-405, 2008.

HEATON, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. **Genetic Programming and Evolvable Machines**, [s. l.], v. 19, n. 1–2, p. 305-307, 2018.

MCCLINTOCK, B. The origin and behavior of mutable loci in maize. **Proceedings of the National Academy of Sciences of the United States of America**, [s. l.], v. 36, n. 6, p. 344-355, 1950.

MITA, P.; BOEKE, J. D. How retrotransposons shape genome regulation. **Current opinion in genetics & development**, [s. l.], v. 37, p. 90–100, 2016.