

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE BIOCÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Explorando Epistasia Adaptativa em Humanos e Outros Primatas

Bruna Oliveira Missaggia

Tese submetida ao Programa de Pós-Graduação
em Genética e Biologia Molecular da UFRGS
como requisito parcial para a obtenção do grau de
Doutora em Genética e Biologia Molecular

Orientadora: Profa. Dra. Maria Cátira Bortolini

Coorientador: Prof. Dr. Márcio Dorn

Porto Alegre, 2024

Esta tese foi realizada no Laboratório de Evolução Humana e Molecular (LEHM) do Departamento de Genética, Instituto de Biociências da Universidade Federal do Rio Grande do Sul, no Laboratório de Laboratório de Bioinformática Estrutural e Biologia Computacional (SBCB) do departamento de informática da Universidade Federal do Rio Grande do Sul e no Laboratory of Computational and Quantitative Biology (LCQB) da Sorbonne Université. Este trabalho foi subsidiado por recursos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A bolsa de doutorado foi concedida por meio do Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM – UFRGS) e através do programa CAPES STIC/AMSUD (Cooperação em Ciência e Tecnologia da Informação e da Comunicação França-América do Sul - CAPES/CDEFI).

Agradecimentos

Agradeço à minha orientadora, Profa. Dra. Maria Cátira Bortolini, pelo empenho com que me ajudou em todas as etapas do desenvolvimento desta pesquisa. A ela também sou grata pelos inúmeros ensinamentos e por toda a generosidade científica e humana que demonstra ao conduzir o LEHM.

Ao meu coorientador, Prof. Dr. Márcio Dorn, por compartilhar seu conhecimento de maneira tão dedicada e inspiradora. Agradeço a ele também por disponibilizar a infraestrutura computacional do SBCB, e pela valiosa oportunidade que me deu ao convidar para fazer meu estágio doutoral na Sorbonne.

À Profa. Dra. Juliana Bernardes, por ter me recebido de forma acolhedora na França e pela orientação nos trabalhos que desenvolvemos por lá.

Agradeço aos colegas do LEHM, do SBCB e do LCQB pela convivência agradável. Sou grata à Dra. Bibiana Fam, ao Dr. Pedro Vargas-Pinilla e ao Lucca Fanti por colaborarem nesta pesquisa.

Aos funcionários e professores do PPGBM, pela competência e ensinamentos. Especialmente ao Elmo Cardoso, por ser sempre tão gentil e solícito, e ao professor Dr. Aldo Mellender Araújo pela orientação durante a graduação.

Aos meus familiares e amigos agradeço o suporte e incentivo sempre. Em especial à minha mãe, pelo apoio, e aos meus irmãos pela cumplicidade. Ao meu pai, *in memoriam*, por tudo dele que vive em mim e naqueles que tiveram o privilégio de conhecê-lo.

À CAPES, pela concessão da bolsa de pesquisa.

RESUMO

A presente tese investigou a relação entre genes e fenótipos adaptativos em humanos e outros primatas, por meio de dois estudos independentes, um de natureza macroevolutiva e o outro em escala microevolutiva. Ambos os trabalhos visam identificar a epistasia adaptativa, utilizando a metáfora da paisagem de aptidão para compreender os padrões observados nos "mapas genótipo-fenótipo adaptativos". O estudo microevolutivo partiu da "hipótese da vitamina D-folato", a qual propõe a radiação ultravioleta como a principal pressão seletiva para a coloração da pele humana. Segundo esta hipótese, não seria apenas por meio da variação na cor de pele que as populações se adaptam aos seus ambientes. Assim, a cor de pele diferente do esperado em nativos americanos não significa necessariamente que essas populações não estejam adaptadas aos seus ambientes. Tais diferenças poderiam ser atribuídas à complexidade da paisagem de aptidão em relação à radiação, o que envolve diferentes vias e a possível coadaptação de alelos entre elas. Ao testar a hipótese, identificamos redes de alelos em duas rotas genéticas relacionadas à adaptação ao ambiente de radiação — coloração da pele e metabolismo da vitamina D — em populações nativas americanas. As possíveis relações epistáticas que diferem entre grupos que vivem em condições ambientais diferentes podem ajudar a explicar a exceção do padrão de variação de cor nessas populações. O estudo macroevolutivo empregou técnicas de Inteligência Artificial para desenvolver o método ProteinPhenotypeInsights (ProPhIn). Este método enfoca a interpretabilidade e visa auxiliar biólogos evolutivos na compreensão das relações entre proteínas candidatas e fenótipos categóricos de interesse em espécies relacionadas. O ProPhIn tem dois objetivos principais: selecionar variações genéticas associadas ao fenótipo e interpretar os padrões gerais dessas variações em relação ao fenótipo e à relação entre as variações selecionadas e as espécies. Ao aplicar o ProPhIn aos dados de 64 espécies de primatas para genes candidatos do sistema oxitocinérgico, identificamos variações genéticas estatisticamente associadas aos fenótipos monogamia social, cuidado biparental e tamanho de ninhada. Além disso, pudemos visualizar os padrões de variação genética e formular hipóteses sobre a relação entre as espécies, as redes de atributos genéticos e os fenótipos na paisagem adaptativa.

ABSTRACT

The present thesis investigated the relationship between genes and adaptive phenotypes in humans and other primates through two independent studies, one of macroevolutionary nature and the other on a microevolutionary scale. Both endeavors aimed to identify adaptive epistasis, utilizing the fitness landscape metaphor to comprehend the patterns observed in adaptive genotype-phenotype maps. The microevolutionary study was grounded on the "vitamin D-folate hypothesis," proposing ultraviolet radiation as the primary selective pressure for human skin coloration. According to this hypothesis, population adaptation to environments cannot be solely explained by skin color variation. Thus, unexpected skin color in Native American populations does not necessarily indicate lack of adaptation to their environments. These differences could be attributed to the complexity of the fitness landscape concerning radiation, which entails considering different pathways of phenotypic adaptation and the potential co-adaptation of alleles among them. Testing this hypothesis, we identified allele networks in two genetic pathways related to radiation environment adaptation - skin coloration and vitamin D metabolism - in Native American populations. Epistatic relationships differing among groups living in different environmental conditions may help explain deviations from the color variation pattern in these populations. The macroevolutionary study employed Artificial Intelligence techniques to develop the ProteinPhenotypeInsights (ProPhIn) method. This method focuses on interpretability and aims to assist evolutionary biologists in understanding the relationships between candidate proteins and categorical phenotypes of interest in related species. ProPhIn has two primary objectives: selecting genetic variations associated with the phenotype and interpreting the general patterns of these variations concerning the phenotype and the relationship between the selected variations and the species. Applying ProPhIn to data from 64 primate species for candidate genes related to the oxytocinergic system, we identified genetic variations statistically associated with social monogamy, biparental care, and litter size phenotypes. Additionally, we visualized patterns of genetic variation and formulated hypotheses about the relationship between species, genetic attribute networks, and phenotypes in the adaptive landscape.

LISTA DE FIGURAS

- Figura 1: Filogenia dos primatas feita a partir de sequências de genoma completo de 50 espécies de primatas. As duas espécies usadas como grupo externo foram o colugo conhecido como lêmure voador de Sunda (*Galeopterus variegatus*- ordem Dermoptera) e o musaranho chinês (*Tupaia belangeri*- ordem Scandentia). A filogenia genômica compreende 14 das 16 famílias de primatas e representa as principais divisões taxonômicas (fonte: Shao et al., 2023 com modificações).....27
- Figura 2: Árvore de decisão da monogamia social com os 8 atributos genéticos conhecidos..... 48
- Figura 3: Distribuição das 64 espécies da amostra nas 14 famílias representadas.. 52
- Figura 4: Filogenia das espécies amostradas com o respectivo fenótipo. A estrela corresponde à monogamia social, o círculo ao cuidado biparental e o quadrado ao parto gemelar frequente. A imagem da árvore filogenética foi gerada a com o iTOL v5 (Interactive Tree Of Life, <https://itol.embl.de/>; Letunic & Bork, 2021) a partir da informação do TimeTree (<https://timetree.org/>; Kumar et al., 2022).....54
- Figura 5: Visão geral sobre a pipeline desenvolvida que é composta por cinco etapas e compreende dois objetivos principais (1) selecionar variações genéticas associadas ao fenótipo e (2) interpretar o que os padrões gerais das variações selecionadas dizem sobre o fenótipo e avaliar a relação entre as variações selecionadas e as espécies..... 57
- Figura 6: Exemplo de arquivos de input utilizados no programa. A) A tabela .csv deve conter pelo menos três colunas, uma com a informação taxonômica, outra com o nome da espécie e a seguinte com um fenótipo categórico binário representado pelos valores zero e um - a classe precisa ser representada por um número real, do tipo ponto flutuante ('float'), ou seja, 0.0 e 1.0. Os nomes compostos das espécies devem ser unidos pelo caractere '_' (ex. *Saimiri boliviensis*). Nos arquivos de alinhamento os nomes das espécies são reconhecidos independentemente do tipo de espaçamento ou da existência de outras informações em volta, portanto a linha de cabeçalho dos arquivos '.fasta' não precisam ser modificadas para entrar no programa..... 62
- Figura7: Plot da distribuição das espécies por família..... 63
- Figura 9: Os gráficos de pizza ajudam a visualizar a distribuição do fenótipo alvo entre as famílias..... 64
- Figura 10: A imagem consiste no plot padrão do scikit-learn de uma árvore de decisão indicando o que representa cada retângulo (nó raiz, interno ou folha) e cujas setas representam os ramos. As cores dos retângulos indicam a classe da maioria das instâncias que estão sendo classificadas em cada parte da árvore (laranja-classe 0, azul- classe 1, branco- metade de cada classe). A árvore acima separa as 64 espécies da amostra em monogâmicas e não-monogâmicas utilizando os rótulos de todas as espécies e todos os atributos. O nó raiz utiliza o atributo L428_V, as 10 espécies que têm o aminoácido V na posição 428 do LNPEP são classificadas como monogâmicas e vão para a primeira folha azul a direita e as demais são direcionadas ao primeiro nó interno a esquerda. As 54 amostras ainda não classificadas são separadas pela presença ou ausência do aminoácido L na posição 177 do receptor de OXT. As espécies que não têm o L nesta posição são classificadas como monogâmicas e vão para a segunda folha em azul a esquerda e as outras seguem para o próximo nó interno onde outra regra de decisão será usada

para separá-las. O algoritmo de árvore de decisão parte do atributo mais relevante, usado no nó raiz, e segue recursivamente até que todas as espécies sejam classificadas em folhas. As folhas podem ser puras, se contém apenas instâncias da mesma classe, ou impuras quando há mistura de classe.....	72
Figura 11: A) A árvore ilustrada é a mesma da figura 10. Os círculos representam as folhas com instâncias da classe 1 (no caso do exemplo, as espécies monogâmicas). O círculo interno, marrom, contém os nomes das espécies que estão na respectiva folha, as outras espécies da classe 1 estão na parte rosa do círculo. B) A imagem representa a árvore de um conjunto de dados com o número reduzido de atributos para ilustrar o plot nos casos em que há folhas impuras. A parte verde dos círculos internos das folhas em que há mistura de classes contém os nomes das espécies da classe 0 que não podem ser separadas daquelas da classe 1 na parte marrom do círculo.....	73
Figura 12. Imagem com os nomes dos atributos pré-selecionados por cada um dos métodos.....	74
Figura 13. A) Gráfico de barras gerado pela função 'analyze_feature_importance_grantham_colors'. O tamanho da barra representa a importância relativa de cada atributo e as cores das barras correspondem a escala do score de Grantham (cujo significado será explicado no tópico seguinte). B) Ilustração do método de pré-seleção multivariado em que são feitos testes de classificação partindo do atributo melhor ranqueado até a correta classificação de todas as espécies da amostra. O gráfico mostra o ganho de acurácia e ao lado quais espécies foram classificadas incorretamente a cada teste.....	76
Figura 14: Proteínas candidatas com as respectivas posições destacadas.....	80
Figura 15: Posições das variações selecionadas com o respectivo aminoácido nas espécies da categoria alvo com as cores determinadas pelo score de Grantham em contraste com a categoria não alvo. O aminoácido mais frequente na posição 362 do receptor de oxitocina em ambas as classes é a A. Como a segunda mais frequente na categoria não alvo (T) é a que difere, as cores do score de grantham ficaram baseadas no contraste do aminoácido da respectiva espécie com o T.....	81
Figura 16: A importância relativa dos atributos para cada árvore de decisão é usada para gerar a importância média. As barras que representam os atributos têm cores que correspondem a Score de Grantham.....	84
Figura 17: A figura ilustra a similaridade entre as espécies <i>Cheirogaleus medius</i> , <i>Ateles geoffroy</i> , <i>Leontopithecus rosalia</i> e <i>Homo sapiens</i> para 5 atributos. A) Tabela com a informação dos 5 atributos para as espécies do exemplo. B) Matriz de similaridade entre as espécies do exemplo. Os valores da diagonal da matriz de similaridade são sempre 1, já que representam a similaridade de uma espécie com ela própria.....	89
Figura 18: A matriz de dissimilaridade é o exato oposto da matriz de similaridade do Simple Matching (Figura 17 B).....	90
Figura 19: A imagem ilustra a os valores de t-SNE em diferentes random state e suas respectivas projeções. É possível observar na tabela que os valores de t-SNE se alteram drasticamente, entretanto as relações de distância entre as espécies são mantidas nas duas projeções.....	91
Figura 20: Mapa de calor da matriz de similaridade entre as espécies baseada no Simple Maching.....	92

Figura 21: No exemplo com um conjunto de dados com o número reduzido de atributos genéticos, podemos observar que algumas espécies se localizam exatamente nas mesmas coordenadas. A forma como é feito o plot permite identificar rapidamente quais espécies são iguais em relação aos atributos usados como entrada. Isso é especialmente relevante quando espécies com fenótipos diferentes se sobrepõem, indicando que os atributos dos genes selecionados não são capazes de distinguir entre os fenótipos para aquelas espécies..... 93

Figura 22: A imagem acima mostra a distribuição das amostras conforme a variação no dado neutro (similaridade das espécies com base nos atributos do citocromo b). Os plots foram feitos a partir das coordenadas do t-SNE que utilizou random state=0. A) Quando utilizamos as famílias como rótulos, é possível observar que o agrupamento é relativamente bom, sem mistura de famílias e com coeficiente de silhueta de 0.43477971731451986. B) Os clusters neste caso são os estados fenotípicos 1 ou 0 (monogâmico 1, não-monogâmico 0). Neste exemplo, quando utilizamos como rótulo a monogamia, observamos que o agrupamento não é tão bom, com coeficiente de silhueta de 0.13353991169361185. Ao comparar as figuras, é possível ver que o dado neutro explica muito melhor a distribuição das famílias do que o fenótipo..... 94

Figura 23: Apesar da alteração da disposição das espécies no plot, o coeficiente de silhueta, medida usada para validação, não sofre grande alteração rodando com diferentes random state..... 95

Figura 24: A figura representa as distribuições das espécies considerando os atributos dos genes candidatos e do dado neutro e os respectivos coeficientes da silhueta para o fenótipo monogamia. As imagens do plot foram geradas com o random state 0. A subtração entre a mediana dos valores de silhueta resulta em um valor positivo de 0.18835151, então podemos constatar que o fenótipo é melhor representado pelos atributos selecionados do que pelo dado neutro..... 97

Figura 25: Matriz de similaridade baseada na co-ocorrência de folha..... 99

Figura 26: Semelhança entre as espécies com relação à co-ocorrência de folha... 100

Figura 27: O gráfico representa a matriz de confusão média dos 6.400 modelos preditivos gerados para testar a taxa de erro das 64 espécies da amostra com relação ao fenótipo monogamia social. As espécies monogâmicas (1) são a classe positiva e as não-monogâmicas (0) são a classe negativa. Na vertical estão reportadas as médias reais e na horizontal as médias preditas. O quadrado interno da parte superior direita com o número 44.37 representa a média de verdadeiros negativos (VN), 0 nos eixos vertical e horizontal. Ou seja, o número médio de espécies que não são monogâmicas que foram preditas corretamente como não monogâmicas (fenótipo 0). O quadrado ao lado, com o número 0.63 representa os falsos positivos (FP), média do número de espécies não monogâmicas (fenótipo 0) que foi predita como monogâmica (fenótipo 1). O quadrado com o número 3.77 representa a média de falso negativos (FN) e o com o número 15.23 a média de verdadeiro positivos (VP)..... 103

Figura 28: Os tamanhos das barras horizontais nos gráficos A e B mostram respectivamente a contagem dos atributos decisivos para o fenótipo monogâmicos e não monogâmicos. A cor das barras, do roxo escuro ao amarelo, representa o percentual de espécies com o mesmo fenótipo que utilizou cada atributo como decisivo. Quanto mais próxima do amarelo a cor da barra, mas espécies com o respectivo fenótipo compartilham o atributo, quanto mais azulado ou próximo do roxo

menos espécies.....	106
Figura 29: A parte de cima da imagem ilustra um exemplo do gráfico de acurácia por iteração até a centésima. A parte de baixo apresenta o resultado das métricas e o percentual de erros por espécie em que aconteceu algum erro de classificação....	108
Figura 30: Os gráficos ilustram a similaridade de espécies monogâmicas com relação aos atributos decisivos, ou seja, aqueles utilizados para separá-las em uma folha pura. A barra lateral mostra o número de atributos decisivos de cada espécie, por exemplo, foram utilizados 24 atributos diferentes para separar a espécie <i>Pithecia pithecia</i> de espécies não monogâmicas. A barra superior apresenta o número de espécies que compartilham a quantidade de atributos da barra inferior.....	109
Figura 31: As barras horizontais nos gráficos A e B mostram a contagem dos atributos decisivos em duas espécies diferentes. A cor das barras, do roxo escuro ao amarelo, representa o percentual de espécies com o mesmo fenótipo que utilizou cada atributo como decisivo. Quanto mais próxima do amarelo a cor da barra, mas espécies com o respectivo fenótipo compartilham o atributo, quanto mais próximo do roxo menos espécies. A) gráfico dos atributos decisivos da espécie da classe alvo (no exemplo, monogamia social) <i>Cercopithecus neglectus</i> . Apenas dois atributos foram usados para separar esta espécie das não monogâmicas semelhantes a ela. Ambos (R177_L_0 e R177_R_1) estão representados por cores escuras, portanto são raros entre os monogâmicos. O atributo decisivo R177_R_1 apresenta a cor mais extrema da escala, indicando que é exclusivo do <i>Cercopithecus neglectus</i> . B) A maior parte dos 14 atributos que foram usados para distinguir a espécie não monogâmica <i>Prolemur simus</i> das monogâmicas semelhantes a ele são frequentes entre os não-mongâmicos, por isso as barras que os representam têm cores claras entre verde e amarelo. Dentre os atributos decisivos do <i>Prolemur simus</i> o 375_H_0 foi o menos frequente entre os não mongâmicos por isso ele está com uma cor mais escura que os outros.....	111
Figura 32: Visualização de quais espécies têm cada conjunto de variações selecionadas como potencialmente importantes. No eixo horizontal estão os atributos. Cada espécie é representada por uma cor diferente, se a espécie tiver a variação, sua respectiva cor será mostrada na posição do atributo.....	113
Figura 34: Um exemplo de rede de atributos que representa o nível de conexão entre eles em relação às espécies da categoria alvo.....	115
Figura 35: Proporção de espécies com o fenótipo alvo e não-alvo (A). Distribuição do fenótipo alvo entre as famílias (B).....	116
Figura 36: Árvore de decisão de classificação das espécies para o fenótipo monogamia social.....	117
Figura 37: Atributos por método de pré-seleção fenótipo monogamia social.....	118
Figura 38: Posições selecionadas na filogenia.....	119
Figura 39: Posições selecionadas dentro das proteínas.....	119
Figura 40: Atributos selecionados ranqueados por importância.....	122
Figura 41: Distribuição das espécies com base nos atributos A) do citocromo b e B) dos genes candidatos selecionados. Em ambos os casos foram utilizadas as famílias como rótulo.....	125
Figura 42: Distribuição das espécies conforme os atributos selecionados.....	127
Figura 43: Distribuição das espécies de acordo com a semelhança por coocorrência de folha.....	129

Figura 44: Atributos decisivos das classes A) monogâmicos sociais e B) não monogâmicos.....	132
Figura 45: Matriz de confusão, métricas de predição e nível de especificidade das espécies por identificação de exceção por percentual de erro.....	134
Figura 46: Gráficos com a contagem e compartilhamento dos atributos decisivos por classe. A) não monogâmicos. B) monogâmicos sociais.....	135
Figura 47: Atributos decisivos das espécies A) <i>Otolemur garnettii</i> e B) <i>Cheirogaleus medius</i>	139
Figura 48: Visualização de quais espécies têm cada conjunto de variações selecionadas como potencialmente importantes. No eixo horizontal estão os atributos. Cada espécie é representada por uma cor diferente, se a espécie tiver a variação, sua respectiva cor será mostrada na posição do atributo.....	143
Figura 49: A rede de espécies que representa o nível de conexão em relação aos atributos selecionados.....	144
Figura 50: A rede de atributos que representa o nível de conexão entre eles em relação às espécies da categoria alvo.....	146
Figura 51: Proporção de espécies com o fenótipo alvo e não-alvo (A). Distribuição do fenótipo alvo entre as famílias (B).....	147
Figura 52: Árvore de decisão de classificação das espécies para o fenótipo cuidado paterno.....	148
Figura 53: Atributos por método de pré-seleção fenótipo cuidado paterno.....	148
Figura 54: Posições selecionadas na filogenia.....	149
Figura 55: Posições selecionadas dentro das proteínas.....	149
Figura 56: Atributos selecionados ranqueados por importância.....	150
Figura 57: Comparação entre a distribuição das espécies conforme a sua semelhança em relação a diferentes critérios: A) Atributos do citocromo b; B) Atributos selecionados dos genes candidatos; C) coocorência de folha.....	151
Figura 58: Contagem de atributos decisivos das classes. Os tamanhos das barras horizontais nos gráficos A e B mostram respectivamente a contagem dos atributos decisivos para o fenótipo cuidado biparental e sem cuidado paterno. A cor das barras, do roxo escuro ao amarelo, representa o percentual de espécies com o mesmo fenótipo que utilizou cada atributo como decisivo. Quanto mais próxima do amarelo a cor da barra, mais espécies com o respectivo fenótipo compartilham o atributo, quanto mais azulado ou próximo do roxo menos espécies.....	152
Figura 59: Matriz de confusão, métricas de predição e nível de especificidade das espécies por identificação de exceção por percentual de erro.....	153
Figura 60: Gráficos com a contagem e compartilhamento dos atributos decisivos por classe. A) espécies com cuidado paterno B) espécies sem cuidado paterno.....	154
Figura 61: Visualização de quais espécies têm cada conjunto de variações selecionadas como potencialmente importantes. No eixo horizontal estão os atributos. Cada espécie é representada por uma cor diferente, se a espécie tiver a variação, sua respectiva cor será mostrada na posição do atributo.....	155
Figura 62: Redes. A) A rede de espécies que representa o nível de conexão em relação aos atributos selecionados. B) A rede de atributos que representa o nível de conexão entre eles em relação às espécies da categoria alvo.....	156

Figura 63: Proporção de espécies com o fenótipo alvo e não-alvo (A). Distribuição do fenótipo alvo entre as famílias (B).....	159
Figura 64: Árvore de decisão de classificação das espécies para o fenótipo tamanho de ninhada.....	160
Figura 65: Atributos por método de pré-seleção fenótipo tamanho de ninhada.....	161
Figura 66: Posições selecionadas na filogenia.....	161
Figura 67: Posições selecionadas dentro das proteínas.....	162
Figura 68: Atributos selecionados ranqueados por importância.....	163
Figura 69: Comparação entre a distribuição das espécies conforme a sua semelhança em relação a diferentes critérios: A) Atributos do citocromo b; B) Atributos selecionados dos genes candidatos; C) coocorrência de folha.....	164
Figura 70: Contagem de atributos decisivos das classes. Os tamanhos das barras horizontais nos gráficos A e B mostram respectivamente a contagem dos atributos decisivos para o fenótipo partos múltiplos e um filhote por ninhada. A cor das barras, do roxo escuro ao amarelo, representa o percentual de espécies com o mesmo fenótipo que utilizou cada atributo como decisivo. Quanto mais próxima do amarelo a cor da barra, mais espécies com o respectivo fenótipo compartilham o atributo, quanto mais azulado ou próximo do roxo menos espécies.....	165
Figura 71: Matriz de confusão, métricas de predição e nível de especificidade das espécies por identificação de exceção por percentual de erro.....	166
Figura 72: Gráficos com a contagem e compartilhamento dos atributos decisivos por classe. A) Mais de um filhote por ninhada/ partos gemelares frequentes. B) Um filhote por ninhada.....	167
Figura 73: Visualização de quais espécies têm cada conjunto de variações selecionadas como potencialmente importantes. No eixo horizontal estão os atributos. Cada espécie é representada por uma cor diferente, se a espécie tiver a variação, sua respectiva cor será mostrada na posição do atributo.....	168
Figura 74: Redes de conexão entre espécies e atributos. A) A rede de espécies que representa o nível de conexão em relação aos atributos selecionados. B) A rede de atributos que representa o nível de conexão entre eles em relação às espécies da categoria alvo.....	169
Figura 75- Correlação entre os fenótipos na amostra.....	173
Figura 76: Relação entre os três fenótipos investigados. Diagrama de Venn com os nomes das espécies.....	174
Figura 77: Distribuição das categorias fenotípicas.....	175
Figura 78: Atributos selecionados para cada um dos fenótipos.....	178
Figura 79: Diagrama de Venn com os atributos selecionados para cada um dos fenótipos.....	178
Figura 80: Lista das posições que compõe os atributos selecionados para cada um dos fenótipos.....	179
Figura 81: Diagrama de Venn com as posições que compõe os atributos selecionados para cada um dos fenótipos.....	179
Figura Suplementar 1: Árvore filogenética com a topologia das espécies e os comprimentos de ramos gerada pelo website TimeTree (https://timetree.org/ ; Kumar et al., 2022).....	226

LISTA DE TABELAS

- Tabela 1: A tabela representa a relação entre um atributo e os aminoácidos de cada posição que a compõe. A coluna a esquerda representa a informação do atributo R225_L+R353_Q+R367_G-R225_F+R367_N. A terceira, quarta e quinta coluna mostram o aminoácido presente em cada nas três posições que estão sendo representadas no atributo..... 68
- Tabela 2: A espécies que têm 1 na coluna do atributo R251_del, tem a deleção na posição 251 do receptor de oxitocina, as que tem o 0 não tem a deleção. As espécies que têm 1 na coluna do atributo R12_G-R12_E+R387_S tem o aminoácido G na posição 12 do receptor de oxitocina, não tem o aminoácido E nesta posição e não tem o aminoácido S na posição 387, as que tem o 0, ao contrário. Todas as espécies que têm o E na posição 18 do receptor de oxitocina não tem o A nesta posição e vice-versa..... 69
- Tabela 3: Exemplo de um pequeno trecho do DataFrame do Citocromo b. As primeiras colunas contém os dados de entrada da tabela .csv (família, espécie, fenótipo), todas as demais representam os atributos genéticos codificados a partir do alinhamento '.translated.fas'..... 70
- Tabela 4: Tabela com os símbolos usados na quantificação da similaridade entre as espécies. Os números 1 e 0 representam o estado em que aparecem os atributos. As letras representam as combinações de estados de atributos, por exemplo, a letra 'a' representa o número de atributos para os quais ambas as espécies têm no estado 1. A letra 'b' representa o número de atributos em que a espécie 1 tem 0 e a espécie 2 tem 1 e assim por diante..... 87
- Tabela suplementar 1: Fenótipos das espécies utilizadas e respectivas referências.... 219
- Tabela suplementar 2: Referência das sequências dos genes LNPEP, OT, OTR e CytB usadas nas análises..... 223

LISTA DE SIGLAS E ABREVIATURAS

Ala⁸OXT - Oxitocina Alanina 8

AVP - Vasopressina

CYTB- Citocromo b

FDR- *False-discovery rate*

GPCRs - Receptores acoplados às proteínas G

IRAP- Aminopeptidase regulada por insulina

Leu⁸OXT- Oxitocina Leucina 8

LNPEP- Leucil-cistinil aminopeptidase

MA- Milhões de anos atrás

mRNA- RNA mensageiro

OXT - Oxitocina

OXTR - Receptor da Oxitocina

PCA- Análise de componentes principais

Phe²OXT - Oxitocina Fenilalanina 2

P-LAP- Leucina aminopeptidase placentária

Pro⁸OXT - Oxitocina Prolina 8

SNPs- *Single-nucleotide polymorphism*

Thr⁸OXT - Oxitocina Treonina 8

t-SNE- *t-distributed Stochastic Neighbor Embedding*

UV- Radiação ultravioleta

UVMED- dose mínima de eritema

Val³Pro⁸OXT - Oxitocina Valina 3 Prolina 8

VDR - Receptor de vitamina D

SUMÁRIO

CAPÍTULO I: INTRODUÇÃO.....	17
CAPÍTULO II: OBJETIVOS GERAIS.....	23
CAPÍTULO III: EM BUSCA DE EPISTASIA ADAPTATIVA- ESTUDO INTERESPECÍFICO (MACROEVOLUTIVO).....	24
3.1. ORDEM PRIMATES.....	24
3.2. FENÓTIPOS: MONOGAMIA SOCIAL, CUIDADO PATERNO E TAMANHO DE NINHADA.....	28
3.3. ESTADO DA ARTE: RELAÇÃO ENTRE OS GENES E OS FENÓTIPOS.....	35
3.4. CONVERGÊNCIA E PARALELISMO EVOLUTIVO.....	45
3.6. OBJETIVOS ESPECÍFICOS.....	51
3.7. MATERIAL E MÉTODOS.....	52
3.7.1. Informações taxonômicas, filogenéticas e fenotípicas da amostra.....	52
3.7.2. Obtenção das sequências e alinhamento.....	56
3.8. RESULTADOS E DISCUSSÃO.....	56
3.8.1. O SISTEMA DE ANÁLISE PROPOSTO: PROTEIN PHENOTYPE INSIGHTS (ProPhIn).....	56
3.8.1.1. VISÃO GERAL DO SISTEMA DE ANÁLISE.....	56
3.8.1.2. DEPENDÊNCIAS DO ProPhIn.....	58
3.8.1.3. LISTA DAS FUNÇÕES DO ProPhIn.....	60
3.8.1.3.1. Funções chamadas pelo usuário.....	60
3.8.1.3.2. Funções auxiliares internas.....	61
3.8.3.4. TUTORIAL:.....	62
PASSO I- Criação dos arquivos de entrada e visualização.....	62
a. Arquivos de entrada (input).....	62
i. Procedimentos para criação do input.....	63
1. Codificação das informações taxonômicas e fenotípicas.....	63
2. Sequências.....	63
b. Visualização.....	64
i. Visualização taxonômica:.....	64
ii. Verificação de balanceamento entre fenótipos e heterogeneidade do desfecho por família.....	64
PASSO II- Criação dos DataFrames e codificação.....	66
a. Criação dos DataFrames.....	66
i. Variação de aminoácidos no alinhamento.....	66
b. Codificação.....	67
i. Nomeação das variantes.....	67
ii. Nomeação dos atributos.....	67
PASSO III- Verificação da capacidade explicativa e pré-seleção de variações genéticas.....	72
a. Visualização da separatividade potencial das classes a partir dos atributos dos genes candidatos.....	72
b. Pré-seleção de atributos.....	75
i. Pré-seleção univariada.....	77

PASSO IV- Seleção de variações genéticas (atributos).....	78
a. Regressão logística filogenética.....	78
b. Visualização das variações.....	79
i. Visualização das posições no contexto da proteína.....	80
ii. Distância de Grantham e classificação de atributos e posições.....	81
a. Ranqueamento das variações pré-selecionadas.....	84
PASSO V- Visualização do padrão geral da variação genética associada ao fenótipo..	85
a. Distribuição espacial das espécies e ganho de informação.....	85
i. Coeficiente de similaridade.....	87
ii. t-SNE.....	90
iii. Plot bidimensional da distribuição das espécies.....	93
iv. Cálculo do coeficiente de Silhueta.....	94
v. Comparação entre coeficientes de silhueta.....	96
b. Co-ocorrência de folha e similaridade entre as espécies.....	99
c. Verificação da capacidade de generalização do modelo.....	102
d. Frequência dos atributos decisivos por classe.....	106
PASSO VI- Interpretação sobre particularidades de espécies.....	107
a. Identificação de exceções.....	108
b. Atributos decisivos por espécie.....	109
ii. Contagem de atributos decisivos por espécie.....	111
c. Análise de redes.....	113
i. Visualização dos atributos selecionados por espécie.....	113
i. Visualização das redes.....	114
1. Visualização das redes de espécies.....	115
2. Visualização das redes dos atributos.....	116
3.8.2. MONOGAMIA SOCIAL.....	116
3.8.2.1 Verificação de balanceamento entre fenótipos e heterogeneidade do desfecho por família:.....	116
3.8.2.2 Visualização da explicabilidade potencial:.....	117
3.8.2.3 Seleção, visualização e ranqueamento dos atributos:.....	118
3.8.2.4 Visualização da distribuição das espécies e comparações de ajuste à taxonomia e ao fenótipo:.....	124
3.8.2.5 Verificação do poder preditivo do modelo e avaliação individual das espécies em relação ao fenótipo:.....	134
3.8.2.6 Visualização dos atributos selecionados nas espécies e redes de espécies e atributos:.....	142
3.8.3 CUIDADO PATERNO INTENSO (OU BIPARENTAL).....	147
3.8.3.1. Verificação de balanceamento entre fenótipos e heterogeneidade do desfecho por família:.....	148
3.8.3.2. Visualização da explicabilidade potencial:.....	148
3.8.3.3. Seleção, visualização e ranqueamento de atributos:.....	149
3.8.3.4. Visualização da distribuição das espécies e comparações de ajuste à taxonomia e ao fenótipo.....	151
3.8.3.5. Verificação do poder preditivo do modelo e avaliação individual	

das espécies em relação ao fenótipo.....	153
3.8.2.6. Visualização dos atributos selecionados nas espécies e redes de espécies e atributos.....	155
3.8.3.7. Discussão cuidado biparental.....	157
3.8.4. TAMANHO DE NINHADA (PARTO GEMELAR).....	159
3.8.4.1. Verificação de balanceamento entre fenótipos e heterogeneidade do desfecho por família:.....	160
3.8.4.2. Visualização da explicabilidade potencial:.....	160
3.8.4.3. Seleção, visualização e ranqueamento dos atributos.....	161
3.8.4.4. Visualização da distribuição das espécies e comparações de ajuste à taxonomia e ao fenótipo.....	164
3.8.4.5. Verificação do poder preditivo do modelo e avaliação individual das espécies em relação ao fenótipo.....	166
3.8.4.6. Visualização dos atributos selecionados nas espécies e redes de espécies e atributos.....	168
3.8.4.7. Discussão tamanho de ninhada (parto gemelar).....	170
3.8.5. DISCUSSÃO GERAL SOBRE OS FENÓTIPOS, AS VARIAÇÕES GENÉTICAS E O MÉTODO DE ANÁLISE PROPOSTO.....	173
3.8.5.1. Fenótipos.....	173
3.8.4.2. Variações genéticas e o sistema de análise proposto.....	179
3.8.6. PERSPECTIVAS.....	187
Capítulo IV: EM BUSCA DE EPISTASIA ADAPTATIVA- ESTUDO POPULACIONAL (MICROEVOLUTIVO).....	188
4.1. PRINCIPAIS RESULTADOS.....	188
4.2. ARTIGO PUBLICADO.....	192
Capítulo V: CONSIDERAÇÕES FINAIS.....	193
6. REFERÊNCIAS.....	199
7. MATERIAL SUPLEMENTAR.....	221

CAPÍTULO I: INTRODUÇÃO

Um dos maiores desafios que as Ciências Biomédicas têm na atualidade é conseguir conectar variações em nível de genoma com aquelas observadas em nível de fenótipos. A complexidade aumenta quando o chamado mapa genótipo-fenótipo (Alberch, 1991) está ligado à aptidão, ou ainda à uma “paisagem adaptativa”, conceito que busca demonstrar como a seleção natural direciona a evolução biológica (De Visser & Krug, 2014; Bank et al., 2022).

A ideia de paisagem de aptidão foi visualizada de forma icônica por Sewall G. Wright ainda na primeira metade do século passado (Wright, 1932). O notável geneticista norte-americano vislumbrou uma paisagem adaptativa como um mapa topográfico com picos e vales, na qual populações evoluem em direção a caminhos que aumentam sua aptidão. Em outras palavras, o modelo tenta explicar como indivíduos de uma população seriam capazes de sobreviver e de se reproduzir ao longo do tempo evolutivo por meio de mutações benéficas, até atingirem um pico de aptidão em determinado contexto. Mudando o contexto e as pressões seletivas, novos picos adaptativos se impõem. Além disso, uma paisagem de aptidão irregular deve possuir múltiplos picos adaptativos que representam soluções genéticas estáveis, mas diferentes, às vezes sub-ótimas, para o desafio de sobreviver (e prosperar ao longo das gerações) em um ambiente específico.

A relação entre espaço genotípico e aptidão veio da ideia de que o valor adaptativo de uma mutação depende do contexto genético em que ela ocorre (Wright, 1931; 1932), já que o autor acreditava na existência de uma epistasia generalizada. O termo epistasia é polissêmico (conforme detalhadamente discutido em Phillips, 1998; 2008), sendo usado para descrever diferentes fenômenos, incluindo a interação funcional entre produtos dos genes, o resultado genético de mutações que atuam na mesma via e o desvio estatístico da ação aditiva (Phillips, 2008). Em um conceito abrangente, a epistasia pode ser definida como qualquer tipo de interação genética em que os efeitos mutacionais dependam do *background* genético (De Visser & Krug, 2014). Existem diversos subtipos ou classes de epistasia dependendo do seu efeito específico, diferentes definições matemáticas e

distintas metodologias para detectar o fenômeno. Domingo (et al., 2019) e Mackay e Anholt (2024) apresentam ótimas revisões sobre estes assuntos.

A epistasia é altamente dependente do equilíbrio entre a eficiência da seleção, a intensidade da deriva e a taxa de mutação (Gros et al., 2009; Sydykova et al., 2020). A variância epistática para a aptidão é a base da teoria da mudança do equilíbrio de Wright, que sugere como ocorrem os movimentos de populações ao longo das paisagens adaptativas considerando os efeitos desses fatores evolutivos (Mackay & Anholt, 2024). Em termos gerais, a epistasia adaptativa é definida como a interação entre alelos de diferentes *loci* em seus efeitos na aptidão (Da Silva et al., 2010). De modo simplificado, ela pode ser entendida como os efeitos não aditivos da interação entre produtos de diferentes genes (intergenômico), ou ainda da interação entre aminoácidos dentro de proteínas (intra-genômico) que resultam em desfechos evolutivos que terão impacto nas populações e, conseqüentemente, na evolução a longo prazo.

A ideia de paisagem adaptativa e relação entre epistasia e fenômenos evolutivos vêm sendo expandida desde os trabalhos pioneiros de Wright. O paleontólogo estadunidense George Simpson e o geneticista russo Theodosius Dobzhansky extrapolaram o conceito de paisagem de aptidão, criada para descrever dinâmicas populacionais (microevolutivas), expandindo a metáfora para o nível macroevolutivo, que representa relações entre espécies (Huneman, 2017). Nesse contexto, cada espécie ocuparia um pico diferente no campo das combinações genéticas. Em paisagens macroevolutivas os vales adaptativos estariam vazios, já que os fenótipos são muito pouco adaptativos. Neste relevo, espécies de nichos ecológicos semelhantes, como o leão e o tigre, ocupariam picos adjacentes, os quais se separariam, por profundos vales, dos picos que ocupam roedores ou primatas, por exemplo (Dobzhansky, 1982). A epistasia também está associada com a especiação, já que geraria as chamadas incompatibilidades de Dobzhansky-Muller. Nesse caso, defende-se que alelos geneticamente neutros ou benéficos na origem de duas populações que divergem passam a ser deletérios em híbridos dessas populações (Orr & Turelli, 2001).

O papel da epistasia adaptativa em contexto macroevolutivo é bem conhecido. Dungan e Chang (2017), por exemplo, levantaram a hipótese de que a

epistasia intramolecular ajudou a conservar as propriedades cinéticas do pigmento visual de pouca luz, a rodopsina, ao mesmo tempo em que permitiu substituições de sintonia espectral com deslocamento para o azul, à medida que os cetáceos se adaptavam ao ambiente aquático. A sinalização é feita através da interação entre pigmento e receptor acoplado à proteína G (GPCR), de modo que os autores descobriram que substituições específicas de aminoácidos nesta molécula promoviam compensações entre os diferentes aspectos funcionais do sistema.

Por outro lado, a epistasia costuma ser negligenciada em nível microevolutivo. Isso acontece porque alguns autores consideram o efeito epistático transitório e pouco importante nesses contextos (Whitlock et al, 1995), além de haver dificuldades técnicas inerentes aos métodos estatísticos que costumam ser usados em dados populacionais de alta dimensionalidade (Pośpiech et al., 2014). Os estudos de GWAS (*Genome-Wide Association Studies*) medem individualmente a associação entre cada variação genética e um fenótipo. Trabalhos de seleção genética populacional costumam assumir que o efeito de uma mutação selecionada é independente do contexto genético em que ela ocorre, e na genética quantitativa clássica a epistasia tende a aparecer como um termo de ruído (Walsh e Lynch, 2018; Tam et al., 2019; Mackay & Anholt, 2024). Porém, cada vez mais os estudos têm mostrado que a epistasia adaptativa abrange todos os níveis biológicos de organização: dentro ou entre genes, dentro de uma população, entre hospedeiro e simbiote, e entre espécies, de modo que o mapa genótipo-fenótipo adaptativo não pode ser completamente entendido sem levar esse fenômeno em consideração (Bank, 2022).

Grande parte do que se conhece sobre o papel da epistasia em nível populacional veio de estudos com organismos modelo, como *Drosophila melanogaster*, *Saccharomyces cerevisiae* e *Escherichia coli*, devido a possibilidade de manipulação experimental (Domingo et al., 2019; Mackay e Anholt, 2024). Diversos estudos mostraram evidências de efeitos fenotípicos diferentes da mesma mutação em diferentes *backgrounds* genéticos, tanto em *S. cerevisiae* quanto em *D. melanogaster* (Gibson & Dworkin, 2004; Dworkin et al., 2009; Yamamoto et al., 2009; Chandler et al., 2017; Galardini et al., 2019). Ainda, uma investigação com mutações introduzidas na enzima isopropilmalato desidrogenase (IMDH) da bactéria

Escherichia coli mostrou sinal de epistasia adaptativa. A IMDH está envolvida na biossíntese de leucina e pode utilizar duas coenzimas alternativas, e o uso de uma delas confere mais aptidão do que o uso da outra. As mutações foram feitas na região que controla o uso das coenzimas por IMDH, alterando a ligação. Os resultados indicam que apesar de cada mutação contribuir de forma aditiva em relação ao uso da coenzima, os efeitos são epistáticos em termos de aptidão (Lunzer et al., 2015). Em estudos com humanos e outras espécies não modelo, as quais não se pode controlar experimentalmente, é muito difícil detectar o fenômeno.

Atualmente é bem conhecido que a epistasia adaptativa determina quais sequências de mutações funcionais consecutivas são viáveis, uma vez que a existência dessas interações limita as rotas disponíveis para que seja atingido um pico adaptativo (Ferretti et al., 2016; Fragata et al., 2019; Østman et al., 2012). Ou seja, uma mudança benéfica para uma característica ou função pode resultar em uma mudança prejudicial para outro traço. Portanto, a seleção natural frequentemente prioriza certas características vantajosas a despeito de outras (Guillaume e Otto, 2012). Esses *trade-offs* (Domingo et al., 2019) são comuns na evolução porque os processos biológicos não ocorrem isoladamente, mas fazem parte dos sistemas altamente integrados e hierárquicos que constituem os organismos vivos. Além disso, os recursos de um organismo são limitados e, muitas vezes, não é possível atingir a adaptação perfeita em todos os fenótipos ao mesmo tempo. Desse modo, a construção dessas paisagens adaptativas requer a seleção das variantes genéticas envolvidas no sinal epistático (Yoo et al., 2020).

Outro fator que pode complexificar o entendimento da relação genótipo-fenótipo em uma paisagem adaptativa é a pleiotropia. Este conceito descreve a capacidade de um único gene ou alelo de influenciar múltiplos fenótipos em um organismo. Em uma definição mais rigorosa, a pleiotropia ocorre quando os efeitos aditivos e/ou de dominância de uma variante polimórfica são diferentes de zero para duas ou mais características (Mackay & Anholt, 2024). Genes do sistema oxitocinérgico seriam casos notáveis. Por exemplo, *OXT* codifica o nonapeptídeo oxitocina que é considerado um neuro-hormônio, uma vez que atua tanto como neurotransmissor no sistema nervoso central, onde é produzido, quanto como hormônio, pois também é liberado na corrente sanguínea e tem um efeito periférico

relacionado a processos fisiológicos, como a contração uterina e ejeção do leite em mamíferos placentários (Zingg e Laporte, 2003; Vargas-Pinilla et al., 2015; Fam et al., 2024). Sua presença no cérebro, por sua vez, modula comportamentos sociais complexos (e.g. formação de casais, cuidados parentais, ver Lucion e Bortolini, 2014; French et al., 2016; Mustoe et al., 2018 e referências lá contidas). A atuação de OXT depende de sua adequada interação com seu receptor preferencial OXTR, caracterizado como uma GPCR, já que atua na cascata de sinalização acoplado a proteínas G (Bockaert e Pin, 1999; Bockaert et al., 2004).

A pleiotropia tem sido considerada um elemento importante para a epistasia (de Visser et al., 2011), na medida que a última necessariamente se torna mais complexa, pois a variação das características também depende frequentemente de interações dos produtos gênicos. Por outro lado, a epistasia também forneceria a variação genética necessária para a evolução da pleiotropia, visto que duas mutações em um gene podem ter menor grau pleiotrópico quando isoladas do que quando juntas (Polster et al., 2016). A epistasia pode reestruturar as correlações genéticas entre características ao combinar padrões de covariação de características favorecidas pela seleção (Jones et al., 2014) e também aliviar as restrições impostas pela pleiotropia (Polster et al., 2016). Desse modo, os efeitos da epistasia podem ser ainda mais importantes do que quando incidem sobre um traço único, alterando os caminhos evolutivos para picos de aptidão mais elevados (Weinreich et al. 2006; Poelwijk et al. 2007; Franke et al. 2011; Polster et al., 2016). Assim, os dois fenômenos estariam correlacionados em um ciclo de influência mútua, moldando paisagens de aptidão e as trajetórias evolutivas das espécies.

Em resumo, a pleiotropia e a epistasia, bem como fenômenos associados (e.g. *trade-offs*), não seriam diferentes daqueles processos encontrados em qualquer outro sistema complexo, nos quais múltiplas partes interagindo devem funcionar juntas para desempenhar as funções adequadamente (Mauro e Ghalambor, 2020). No caso dos organismos vivos, o resultado é sua extraordinária capacidade de se adaptar e evoluir (evolvabilidade).

Incorporar tais fenômenos em estudos evolutivos é difícil devido à complexidade das interações multiloci, resultando em desafios teóricos e estatísticos notáveis (Mackay 2014; Bank, 2022). Desse modo, buscamos, através de dois

trabalhos, responder como um gene (ou um grupo deles) e suas variantes se conectam com fenótipos adaptativos em humanos e outros primatas (“mapa genótipo-fenótipo adaptativo”).

Em um dos estudos que fazem parte de presente tese, e já publicado, utilizamos programa de análise conhecido, mas desenvolvemos uma abordagem estatística própria, para identificar a conexão (redes) de alelos de duas rotas genéticas potencialmente correlacionadas (coloração da pele e metabolismo da vitamina D), considerando populações nativas americanas que vivem em diferentes latitudes e altitudes e com diferentes hábitos de dieta (Missaggia et al., 2020).

O outro estudo envolve uma abordagem macroevolutiva, o qual foi desenvolvido utilizando técnicas de Inteligência Artificial, a partir de um método inovador denominado de *ProteinPhenotypeInsights* (ou *ProPhIn*), capaz de identificar os aminoácidos táxon-específicos em genes pleiotrópicos do sistema oxitocinérgico que potencialmente fariam parte do repertório genético epistático por trás de alguns fenótipos adaptativos, incluindo alguns relativamente raros em mamíferos, tais como monogamia social, intenso cuidado paterno (machos cuidadores) e parto gemelar (Fernandez-Duque et al., 2009; Lukas e Clutton-Brock, 2012; Stockley e Hobson, 2016).