MAGYAR AGRÁR- ÉS
ÉLETTUDOMÁNYI EGYETEM

Development of mitochondrial genomics and proteomics methods

Thesis of doctoral (PhD) dissertation

Bálint Biró

Gödöllő

2022

The doctoral school's

name: Animal Biotechnology and Animal Sciences

discipline: Animal sciences

head: Professor Dr. Miklós Mézes
D.V.M., member of the HAS
Hungarian University of Agriculture and Life Sciences, Szent István Campus, Institute of Physiology and Animal Nutrition, Department of Nutritional Safety

Supervisor(s): Dr. Orsolya Ivett Hoffmann
Senior Research Fellow, Ph.D.
Hungarian University of Agriculture and Life Sciences, Szent István Campus, Institute of Genetics and Biotechnology, Department of Animal Biotechnology

....
Doctoral School's approval
Dr. Miklós Mézes
member of the HAS

....
Supervisor's approval
Dr. Orsolya Ivett Hoffmann
Senior Research Fellow

1 Preliminary information and objectives

1.1 Preliminary information

In the second half of the XXth century, the biggest challenge in life sciences was the absence of proper methods for the deep analysis of biological systems. In this period, several ”game changer” techniques were developed. These new methods like sequencing, X-ray crystallography, NMR etc. forced the scientists to master the required skills to be able to routinely perform them (Gilbert, 1991).

The mentioned methods were absolutely cutting edge technologies back in the days, however due to the complexity, low throughput and time consuming pattern of their usage, the widespread application of them were not possible.

The millenium is considered as a turning point since the scientific community achieved a technological phase where the automation of sequencing became possible (Pauwels et al., 1995). As a consequence of this achievement, huge amount of biological sequence data has been accumulated (O’Leary et al., 2016; “UniProt: the universal protein knowledgebase in 2021”, 2021). Due to this growth in the quantity of biological sequences, the data science attitude appeared in biology resulted in a paradigm shift (D’Argenio, 2018; Pal et al., 2020). However the number of functionally or structurally characterized nucleic acid sequences lags behind the number of raw sequence data (Salzberg, 2019). The same pattern applies to proteins since there is a huge gap between the number of sequenced proteins and the number of characterized proteins too (Schwede, 2013). Hence one of the most recent problems in nowadays biology is to try to annotate raw sequence data. The truth is that achieving this goal -annotate sequences-, purely with wet lab experiments is nearly impossible. Although due to technological achievements, several computational techniques (machine learning based methods) have been developed. With these techniques one is able to extract meaningful patterns from highly complex data structures.

During my doctoral research I was involved in two distinct biological disciplines.

In the first one, nuclear mitochondrial sequences (Lopez et al., 1994)(NU-MTs) were investigated. NUMTs were described in several cancer types where tumor suppressors or oncogenes were affected by NUMT integrations (Ju et al., 2015; Singh et al., 2017; Srinivasainagendra et al., 2017; Palodhi et al., 2020; Wei et al., 2022). Other than tumor biological usage, NUMTs have important applications in phylogenetics (Ko et al., 2015; Nacer & do Amaral, 2017) and forensic studies (Marshall & Parson, 2021; Cortes-Figueiredo et al., 2021). Based on recent findings, NUMTs are also important when it comes to nuclear OFF-target cleavages during mitochondrial targeted genome editing (Lei et al., 2022).

NUMTs were characterized with a wide range of methods in several species including human (Dayama et al., 2014), dog (Verscheure et al., 2015), cat (Lopez et al., 1994) etc. NUMTs have not been described in the rabbit genome yet, even though, several studies show that in some cases, rabbit is a better disease model than rodents or primates (Esteves et al., 2018; Fan et al., 2018; Matsuhisa et al., 2020; Fan et al., 2021). As stated, NUMTs are well known in many species, however high throughput NUMT annotation studies are focusing on small, isolated taxonomical units (G. Zhang et al., 2021; Calabrese et al., 2017; Tsuji et al., 2012). Furthermore these studies are not using a community approved, standardised framework. Therefore the results of these isolated studies without uniform methods are poorly comparable.

The other biological discipline that we were woking on is the evaluation of some proteomics related, machine learning based algorithms. In this section the complementarity of secondary structure, solvent accessibility and nucleic acid interaction were measured on the total human proteome (McGuffin et al., 2000; Faraggi et al., 2014; Yan & Kurgan, 2017). The mentioned features of proteins are important in parasite-host interactions (Kruglikov et al., 2021), protein folding (Savojardo et al., 2021), basic research (Cozzolino et al., 2021) etc. As mentioned previously, the traditional, wet lab approach for the high throughput determination of those features would not be possible and so this is where ML based techniques come in handy. ML is a collective term referring to the implementation and testing of models that are based on existing data and are able to perform recognition, classification and estimation tasks (Tarca et al., 2007). The performance of those models are need to be evaluated. The most widespread approach to evaluate the performance of a model is to try to predict the labels of an unknown dataset (aka test set). The training and testing of ML models are usually performed on small and somehow preselected datasets. The problem with these general and non standardized datasets is that different models achieve different performances on different datasets.

In my doctoral dissertation, NUMT biology and proteomics is linked together by data science. In the NUMT biology module, full life cycle of an ML model is done from feature selection, through hyperparameter optimisation to performance evaluation. In the proteomics module, performance evaluations of three ML models were performed on the human proteome.

1.2 Objectives

1.2.1 NUMT biology

Our main objective in this module was to create an algorithm to detect the NUMTs that are integrated into the rabbit genome. Our next goal was to extend the previously defined ”numt mining algorithm” to all NCBI mammalian genomes.

1.2.2 Proteomics

In the proteomics module our goal was to investigate the residue level complementarity of secondary structure, solvent accessibility and nucleic acid interaction predictions on the human proteome.

2 Materials and methods

2.1 Statistical analysis and machine learning

Statistical analysis was performed with Python's Scipy (version number: 1.6.2) and Numpy (version number: 1.20.3) libraries, while ML models were implemented in scikit-learn (version number: 0.24.2) and umap (version number: 0.5.3) libraries (Virtanen et al., 2021; Harris et al., 2020; Pedregosa et al., 2011; McInnes et al., 2018).

2.2 NUMT biology

Rabbit genome (OryCun2.0) was acquired from the Ensembl database. For the extended NUMT mining workflow, all the genomes and taxonomical data were downloaded from the NCBI database. The newest assemblies were used.

LASTAL (version number: 1219) was used to perform sequence alignment between nuclear and corresponding mitochondrial genomes with the following settings: match=1, mismatch=-1, gap open penalty=7, gap extension penalty=1 (Kiełbasa et al., 2011).

To classify NUMTs and random sequences, RBF-Kernel SVM was trained. k-times cross validation (k=3) was performed to evaluate the model's performance. To avoid data-leakage, input matrices were normalised in every cross validation iteration with the min-max procedure.

The genetic distance of NUMTs and the corresponding mitochondrial sequences were evaluated with the modified Kimura2 parameter which tolerates alignment gaps.

5 kb flanking sequences were extracted with SAMTOOLS (version number: 1.6) and were analysed with RepeatMasker (version number: 4.1.2-p1) (Li et al., 2009).

Phylogenetic analysis and visualisation were performed in R system's ape (version number: 5.6-2) and ggtree (version number: 3.2.1) packages respectively (Paradis & Schliep, 2019; Yu, 2020).

2.3 Proteomics

The human proteome was acquired from the UniProt database ("UniProt: the universal protein knowledgebase in 2021", 2021).

Peptides that were shorter than 30 amino acids were excluded from the downstream analysis. In cases when more PDB chains were corresponding to the same part of a UniProt sequence, the shorter chains were discarded. If more structures were present for the same part of a UniProt sequence, the one with the best resolution was involved in further analysis. These filtering steps resulted in 5 133 UniProt sequences and the corresponding 6 417 PDB structures.

SA and secondary structure data were extracted from PDB files using the DSSP algorithm (Kabsch & Sander, 1983).

BioLip database served as the ground truth data source for nucleic acid binding residues (Yang et al., 2012). The mapping of BioLip annotation to the UniProt sequences resulted in 3 577 DNA binding and 2 368 RNA binding residues in 175 and 106 proteins respectively.

The performance evaluation of the prediction of nucleic acid binding residues was performed with ROC analysis.

The performance of the prediction of secondary structure elements were evaluated with the Segment Overlap (SOV) score.

3 Results and discussion

3.1 NUMT biology

To be able to differentiate NUMTs and random sequences, RBF-kernel SVM model was trained to perform a classification task. The SVM input features were genomic position, genomic length, the GC content of a given sequence and the GC contents of the flanking regions. The SVM model classified the NUMTs and random sequences with approximately 0.7 accuracy in the cross validation (Figure 1.) and outperformed the ”dummy classifier” in every complexity stage. What is more, this model was able to differentiate whether the given NUMT is located on scaffold or chromosome (Equation. 3.1). The model has the best accuracy in the 10^{-16}-10^{-14} γ-parameter range (Figure 1./a). The size of the training dataset proved to be effective (Figure 1./b).

$$CM = \begin{bmatrix} 21 & 0 & 0 & 7 \\ 0 & 13 & 2 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 26 \end{bmatrix} \tag{Equation. 3.1}$$

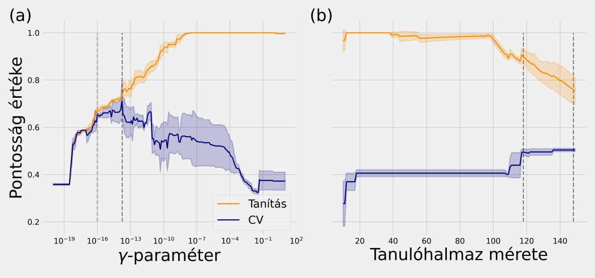

Figure 1: Validation (a) and learning (b) curve of RBF-kernel SVM for classifying NUMTs and random sequences.

Based on the input features, NUMTs clustered together corresponding to their taxonomical orders (Figure 2./a).

The distribution of relative NUMT sizes (compared to the corresponding mitochondrion size) were skewed since shorter NUMTs were overrepresented (Figure 2./b). This tendency seems to be significant (p<0.05). The interquartile range of the absolute NUMT sizes was under 600 bp, while the median was well under 250 bp (Figure 2./b). In case of three species (beluga, bottlenose dolphin and american beaver) we observed greater than 1.0 relative NUMT size, which means that the whole mitochondrion was inserted into the corresponding nuclear genome as s NUMT.

We found out that the in the case of a linear mitochondrial genome, the terminal nucleotides are proned to be involved in the process of NUMTogenesis. (p<0.05) (Figure 2./g).

When we compared the GC ratios of NUMTs to gDNA’s GC ratio, we got a value smaller than 1.0. However, when we comapred the NUMTs’ GC ratios to mtDNA’s GC ratio we got 1.0 (Figure 2./e).

In the 5 kb flanking regions of NUMTs we found several repetitive elemet classes that showed higher frequencies near NUMTs (Figure 2./d). Those repetitive elements were DNA/hAT-Charlie, Simple repeat, LTR/ERV1, LINE/L1, LTR/ERVL-MaLR, DNA/TcMar-Tigger, LTR/ERVL, LINE/L2, SINE/MIR and SINE/Alu.

There is a weak positive relationship (0.378, p<0.0001) between NUMT count and genome size. While we found a weak negative relationship (-0.42, p<0.001) between the cumulated NUMT sizes and genome size (Figure 2./c,f).

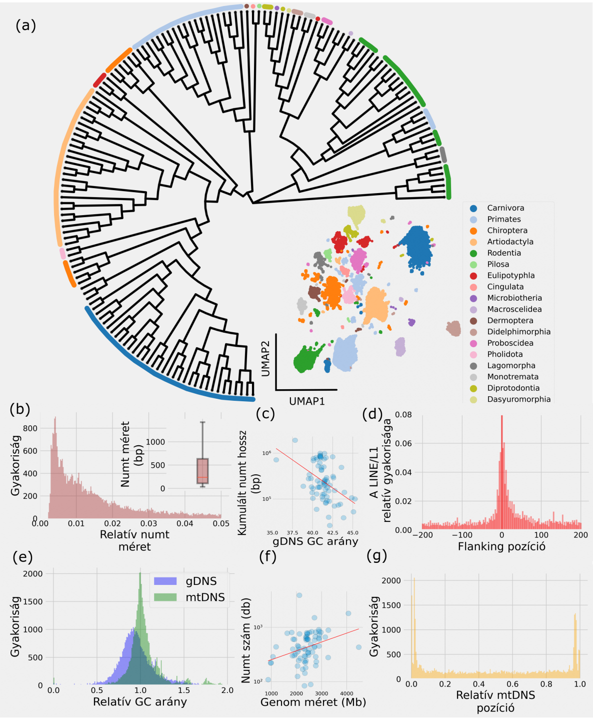

Figure 2: NUMTs of the NCBI mammalian genomes.
UMAP clusters of NUMTs and the phylogenetic tree of the corresponding taxonomical orders (a). Distribution of relative and absolute NUMT sizes (b). Relationships between genome sizes versus cumulative NUMT sizes (c) and genome sizes versus NUMT counts (f). Frequency of LINE/L1 in the 200 bp flanking region of NUMTs (d). Relative GC ratio of NUMT's compared to gDNA and mtDNA (e). Relative positions of NUMTs (g).

3.2 Proteomics

Based on our experiments, residues that are in coil (C) secondary structure have the highest, while residues that are in sheet (E) secondary structure have the lowest relative solvent accessibility-SA (aka most solvent exposed and least solvent exposed respectively). Helical residues (H) are intermediate from the solvent accessibility point of view. Those observation are significant ($p<10^{-5}$) no matter how data sources of solvent accessibility and secondary structure are combined together (Figure 3.).

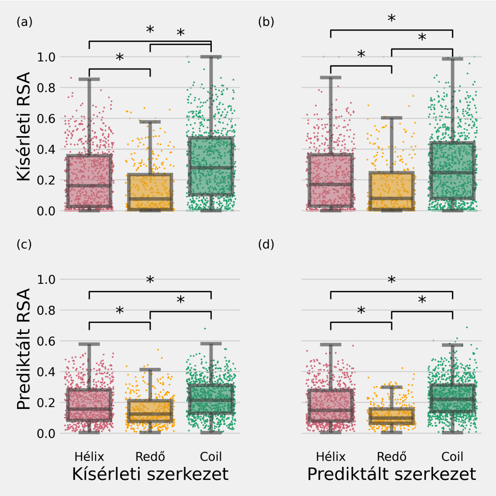

Figure 3: The relationships of SA and secondary structure.
In the upper row (a,b) RSA values are experimentally derived while those data are predicted in the lower row (c,d). In the first column (a,c) secondary structure information is experimentally derived while in the second column (b,d) this data is predicted. * shows significant ($p<10^{-5}$) results.

The performance of nucleic acid binding prediction of DRNApred was evaluated with ROC analysis. AUC of DNA binding was 0.69, while AUC of RNA binding was 0.56 (Figure 4.).

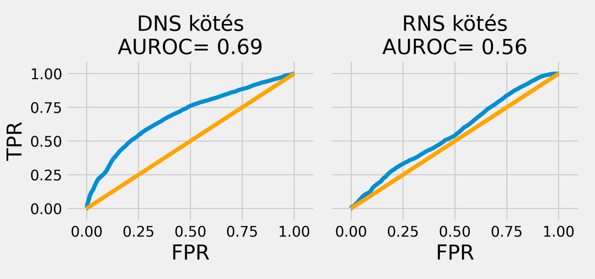

Figure 4: ROC analysis of nucleic acid binding predictions of DRNApred.

During the evaluation of the relationship between nucleic acid binding and SA, we found out that the residues that bind nucleic acids have higher SA while the residues that do not bind nucleic acids have lower SA. i.e. more solvent exposed and less solvent exposed respectively (Figure 5.). The previously mentioned relationships are significant ($p<10^{-5}$) no matter how we combine the data sources of nucleic acid binding and SA. The same applies to RNA binding, however in the thesis pamphlet only the DNA binding results are shown.

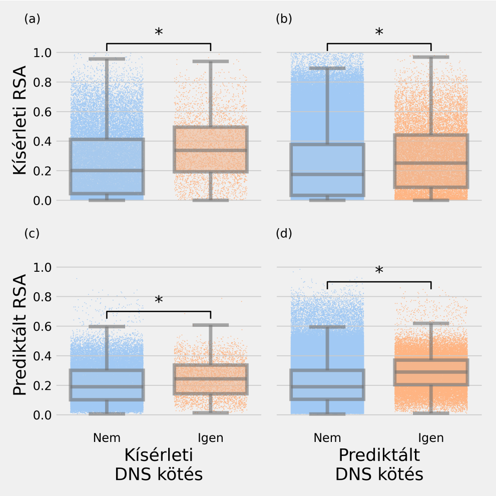

Figure 5: SA and DNA binding.
Upper row (a,b) contains experimentally derived RSA values, while the lower row (c,d) contains predicted RSA values. In the first column (a,c), DNA binding data is experimentally determined while in the second column (b,d), DNA binding is predicted. * stands for significant (p<10^{-5}) results.

4 Conclusions

4.1 NUMT biology

In the mitochondrial genomics part, nuclear sequences with mitochondrial origins aka NUMTs were scrutinised. NUMTs have strong influence on cancer biology research and they have important applications in the fields of phylogeny and forensic studies also. In my dissertation, firstly the NUMTs of the rabbit genomes were characterized. In this part, the genomic coordinates of the NUMTs were defined and so based on these coordinates we were able to identify intragenic NUMTs. By the investigation of the GC content of the flanking regions of NUMTs, it turned out that genomic surrounding of NUMTs have significantly altered, namely the flankings had lower GC content than the rest of the genome. Moreover, we proved that the frequency of several repetitive elements are persistent near the NUMTs. Then, based on the features on NUMTs a SVM was taught to be able to differentiate NUMTs from random sequences. After all, the previously described workflow has been scaled up to all mammalian genomes of NCBI and so a NUMT mining pipeline has been established. The NUMTs from nearly 150 genomes showed distinguishing features based on taxonomy data. The most important result of this NUMT mining workflow is that several repetitive elements displayed elevated frequency in the flanking regions of NUMTs. New scientific finding is that several previously uncharacterised genomes have been investigated from the NUMTogenesis point of view with a uniform methodology. The mitochondrial genomic part of my doctoral dissertation was performed at the Hungarian University of Agriculture and Life sciences, Institute of Genetics and Biotechnology, Department of Animal Biotechnology under the supervision of Dr Orsolya Ivett Hoffmann.

4.2 Proteomics

In the proteomics part of my dissertation, the residue level complementarity of protein structure and function predictors were evaluated on the human proteome. The significance of this topic is that several artificial intelligence based predictors have been published recently which inputs are protein sequences. However the predictive performance of these models are evaluated on somehow preprocessed, prefiltered datasets (subcellular localisation, defined function etc.) even though their inputs are the same sequences. In this section we investigated residue level solvent accessibility, secondary structure and interactibility with nucleic acids. These features are extremely important at different steps of drug design. We proved that state-of-the-art predictive models successfully reproduce the motifs that are present in the experimental dataset. Additionally we found these motifs even when different data sources were combined arbitrarily. This kind of evaluation of predictive methods can help to achieve elevated predictive performance. Moreover this evaluation method is able to facilitate feature engineering/extraction. The proteomics related part of my dissertation was done at the Virginia Commonwealth University, College of Engineering, Department of Computer Science in the structural bioinformatics laboratory of Dr Lukasz Kurgan.

5 New scientific results

1. First annotation of the rabbit genome from the NUMTogenesis point of view
2. We proved the altered GC content of NUMTs in the rabbit genome
3. Some repetitive elements were described that show higher frequencies near NUMTs
4. All the NCBI mammalian genomes were annotated from the NUMTogenesis point of view
5. We proved the complementarity of residue level SA, secondary structure and nucleic acid interaction predictions with experimental data on the human dataset
6. With complementarity checking, we developed a workflow for performance evaluation

6 Published articles in the topic of the dissertation

6.1 Publications with impact factor

- Biró, B., Zhao, B. and Kurgan, L. (2022). Complementarity of the residue-level protein function and structure predictions in human proteins. Computational and structural biotechnology journal, 20, 2223-2234. D1 Biofizika, IF: 7.271
- Biró, B., Gál, Z., Schiavo, G., Ribari, A., Utzeri, V. J., Brookman, M., ... and Hoffmann, O. I. (2022). Nuclear mitochondrial DNA sequences in the rabbit genome. Mitochondrion, 66, 1-6. Q2 Sejtbiológia, IF: 4.35

6.2 Publications without impact factor

- Biró, B., Gál, Z., Brookman, M. and Hoffmann, O. I. (2022). Patterns of NUMTogenesis in sixteen different mice strains. bioRxiv.

References

Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, *20*(4), 477–486.

Ball, P. (2017). Water is an active matrix of life for cell and molecular biology. *Proceedings of the National Academy of Sciences*, *114*(51), 13327–13335.

Calabrese, F., Balacco, D., Preste, R., Diroma, M., Forino, R., Ventura, M., & Attimonelli, M. (2017). Numts colonization in mammalian genomes. *Scientific reports*, *7*(1), 1–10.

Cortes-Figueiredo, F., Carvalho, F. S., Fonseca, A. C., Paul, F., Ferro, J. M., Schönherr, S., … Morais, V. A. (2021). From forensics to clinical research: expanding the variant calling pipeline for the precision id mtdna whole genome panel. *International journal of molecular sciences*, *22*(21), 12031.

Cozzolino, F., Iacobucci, I., Monaco, V., & Monti, M. (2021). Protein–dna/rna interactions: An overview of investigation methods in the-omics era. *Journal of Proteome Research*, *20*(6), 3018–3030.

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic acids research*, *42*(20), 12640–12649.

D'Argenio, V. (2018). The high-throughput analyses era: are we ready for the data struggle? *High-throughput*, *7*(1), 8.

Esteves, P. J., Abrantes, J., Baldauf, H.-M., BenMohamed, L., Chen, Y., Christensen, N., … others (2018). The wide utility of rabbits as models of human diseases. *Experimental & molecular medicine*, *50*(5), 1–10.

Fan, J., Chen, Y., Yan, H., Niimi, M., Wang, Y., & Liang, J. (2018). Principles and applications of rabbit models for atherosclerosis research. *Journal of atherosclerosis and thrombosis*, *25*(3), 213–220.

Fan, J., Wang, Y., & Chen, Y. E. (2021). Genetically modified rabbits for cardiovascular research. *Frontiers in Genetics*, *12*, 14.

Faraggi, E., Zhou, Y., & Kloczkowski, A. (2014). Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins: Structure, Function, and Bioinformatics*, *82*(11), 3170–3176.

Fujiwara, K., Toda, H., & Ikeguchi, M. (2012). Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC structural biology*, *12*(1), 1–15.

Gilbert, W. (1991). Towards a paradigm shift in biology. *Nature*, *349*(6305), 99.

Haque, M. M., & Bayford, R. (2019). *Protein misfolding thermodynamics* (Vol. 10) (No. 10). ACS Publications.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … others (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Ilyina, E., Roongta, V., & Mayo, K. H. (1997). Designing water soluble β-sheet peptides with compact structure. In *Techniques in protein chemistry* (Vol. 8, pp. 797–808). Elsevier.

Ju, Y. S., Tubio, J. M., Mifsud, W., Fu, B., Davies, H. R., Ramakrishna, M., … others (2015). Frequent somatic transfer of mitochondrial dna into the nuclear genome of human cancer cells. *Genome research*, *25*(6), 814–824.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, *22*(12), 2577–2637.

Kalinowska, B., Banach, M., Wiśniowski, Z., Konieczny, L., & Roterman, I. (2017). Is the hydrophobic core a universal structural element in proteins? *Journal of Molecular Modeling*, *23*(7), 1–16.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research*, *21*(3), 487–493.

Ko, Y.-J., Yang, E. C., Lee, J.-H., Lee, K. W., Jeong, J.-Y., Park, K., … Yim, H.-S. (2015). Characterization of cetacean numt and its application into cetacean phylogeny. *Genes & Genomics*, *37*(12), 1061–1071.

Kruglikov, A., Rakesh, M., Wei, Y., & Xia, X. (2021). Applications of protein secondary structure algorithms in sars-cov-2 research. *Journal of Proteome Research*, *20*(3), 1457–1463.

Lei, Z., Meng, H., Liu, L., Zhao, H., Rao, X., Yan, Y., … Yi, C. (2022). Mitochondrial base editor induces substantial nuclear off-target mutations. *Nature*, 1–1.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078–2079.

Lins, L., Thomas, A., & Brasseur, R. (2003). Analysis of accessible surface of residues in proteins. *Protein science*, *12*(7), 1406–1417.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., & O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial dna to the nuclear genome of the domestic cat. *Journal of molecular evolution*, *39*(2), 174–190.

Marshall, C., & Parson, W. (2021). Interpreting numts in forensic genetics: Seeing the forest for the trees. *Forensic Science International: Genetics*, *53*, 102497.

Matsuhisa, F., Kitajima, S., Nishijima, K., Akiyoshi, T., Morimoto, M., & Fan, J. (2020). Transgenic rabbit models: Now and the future. *Applied Sciences*, *10*(21), 7416.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The psipred protein structure prediction server. *Bioinformatics*, *16*(4), 404–405.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Miao, Z., & Westhof, E. (2015). Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic acids research*, *43*(11), 5340–5351.

Mukherjee, S., & Bahadur, R. P. (2018). An account of solvent accessibility in protein-rna recognition. *Scientific reports*, *8*(1), 1–13.

Nacer, D. F., & do Amaral, F. R. (2017). Striking pseudogenization in avian phylogenetics: numts are large and common in falcons. *Molecular phylogenetics and evolution*, *115*, 1–6.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … others (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745.

Pal, S., Mondal, S., Das, G., Khatua, S., & Ghosh, Z. (2020). Big data in biology: The hope and present-day challenges in it. *Gene Reports*, *21*, 100869.

Palodhi, A., Singla, T., & Maitra, A. (2020). Profiling of numts in gingivobuccal oral cancer. *bioRxiv*.

Pan, Y., Zhou, S., & Guan, J. (2020). Computationally identifying hot spots in protein-dna binding interfaces using an ensemble approach. *BMC bioinformatics*, *21*(13), 1–16.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics*, *35*(3), 526–528.

Pauwels, R., Azijn, H., de Béthune, M.-P., Claeys, C., & Hertogs, K. (1995). Automated techniques in biotechnology. *Current Opinion in Biotechnology*, *6*(1), 111–117.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Salzberg, S. L. (2019). *Next-generation genome annotation: we still struggle to get it right* (Vol. 20) (No. 1). BioMed Central.

Savojardo, C., Manfredi, M., Martelli, P. L., & Casadio, R. (2021). Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Frontiers in molecular biosciences*, *7*, 460.

Schwede, T. (2013). Protein modeling: what happened to the "protein structure gap"? *Structure*, *21*(9), 1531–1540.

Singh, K. K., Choudhury, A. R., & Tiwari, H. K. (2017). Numtogenesis as a mechanism for development of cancer. In *Seminars in cancer biology* (Vol. 47, pp. 101–109).

Srinivasainagendra, V., Sandel, M. W., Singh, B., Sundaresan, A., Mooga, V. P., Bajpai, P., … Singh, K. K. (2017). Migration of mitochondrial dna in the nuclear genome of colorectal adenocarcinoma. *Genome medicine*, *9*(1), 1–15.

Tang, Y., Liu, D., Wang, Z., Wen, T., & Deng, L. (2017). A boosting approach for prediction of protein-rna binding residues. *BMC bioinformatics*, *18*(13), 47–58.

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology*, *3*(6), e116.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, *8*(11), e80635.

Tsuji, J., Frith, M. C., Tomii, K., & Horton, P. (2012). Mammalian numt insertion is non-random. *Nucleic acids research*, *40*(18), 9073–9088.

Uniprot: the universal protein knowledgebase in 2021. (2021). *Nucleic acids research*, *49*(D1), D480–D489.

Verscheure, S., Backeljau, T., & Desmyter, S. (2015). In silico discovery of a nearly complete mitochondrial genome numt in the dog (canis lupus familiaris) nuclear genome. *Genetica*, *143*(4), 453–458.

Virtanen, P., Gommers, R., Burovski, E., Oliphant, T. E., Weckesser, W., Cournapeau, D., … others (2021). scipy/scipy: Scipy 1.6. 3. *Zenodo*.

Wang, J.-X., Liu, J., Miao, Y.-H., Huang, D.-W., & Xiao, J.-H. (2020). Tracking the distribution and burst of nuclear mitochondrial dna sequences (numts) in fig wasp genomes. *Insects*, *11*(10), 680.

Wei, W., Schon, K. R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., … Chinnery, P. F. (2022). Nuclear-embedded mitochondrial dna sequences in 66,083 human genomes. *Nature*, *611*(7934), 105–114.

Yan, J., & Kurgan, L. (2017). Drnapred, fast sequence-based method that accurately predicts and discriminates dna-and rna-binding residues. *Nucleic acids research*, *45*(10), e84–e84.

Yang, J., Roy, A., & Zhang, Y. (2012). Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, *41*(D1), D1096–D1103.

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current protocols in bioinformatics*, *69*(1), e96.

Zhang, G., Geng, D., Guo, Q., Liu, W., Li, S., Gao, W., … others (2021). Genomic landscape of mitochondrial dna insertions in 23 bat genomes: characteristics, loci, phylogeny, and polymorphism. *Integrative Zoology*.

Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., & Kurgan, L. (2009). On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, *76*(3), 617–636.

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., & Kurgan, L. (2010). Analysis and prediction of rna-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Current Protein and Peptide Science*, *11*(7), 609–628.

Zhu, Z.-Y., & Blundell, T. L. (1996). The use of amino acid patterns of classified helices and strands in secondary structure prediction. *Journal of molecular biology*, *260*(2), 261–276.