Identifying sequence mutations responsible for ligand binding domain changes

Georgia Wyldbore

43235779

Supervisor: Gabriel Foley

Literature Review

Nuclear receptors (NRs) are a family of 48 transcription factors, characterised through their conserved structural domains, which consist of a highly conserved DNA binding domain, and a somewhat less conserved ligand binding domain or pocket (D'Arrigo et al., 2022; de Vera, 2018; Mazaira et al., 2018; Shulman et al., 2004; Weikum et al., 2018). These transcription factors are responsible for driving recruitment of transcriptional co-regulators, and through this controlling the activation and repression of a variety of target genes. As such, they regulate a wide range of physiological processes including inflammation, reproduction, circadian rhythm and metabolism, among others (Weikum et al., 2018). Because of their important role in so many biological systems, changes in the signalling pathways of NRs can conttibute to disease states such as cancer, diabetes, obesity and Parkinsons, making NRs a valuable target for development of therapeutic drugs (de Vera et al., 2019; Mazaira et al., 2018; Munoz-Tello et al., 2020; Rodriguez-Calvo et al., 2017; Shulman et al., 2004). Indeed, approximately ten to twenty percent of current FDA-approved drugs target NRs, with many more in development (de Vera, 2018; Weikum et al., 2018). As such, improving the knowledge base surrounding NRs may assist in the development of future therapies.

The canonical process through which NRs base their biological function is through the binding of a ligand, which – depending on the specific NR – allows the stabilisation or displacement of a short alpha-helix segment within the structure (D'Arrigo et al., 2022; Munoz-Tello et al., 2020; Weikum et al., 2018). NRs appear to exist in somewhat of an equilibrium between conformational states, with a ligand inducing a transition towards a more ordered state. However, this state instability can make determining exact structure of NRs difficult (D'Arrigo et al., 2022). NRs are often regulated endogenously by small, lipophilic ligands such as steroids, but between ten and twenty remain which are referred to as ‘orphans’, where an in vivo ligand has not yet been identified (de Vera, 2018; Weikum et al., 2018). The literature varies as to exactly which NRs are deemed orphans, as it appears some publications class orphans as ‘adopted’ once any endogenously occurring ligand can be found to bind in vitro, and some consider them adopted if a synthetic ligand has been identified. Additionally, de Vera (2018) acknowledges that it is unclear whether some of the adopted NRs are actually regulated by ligands, despite having the ability to bind them.

The 48 NRs have been further classified into seven sub-families, numbered 0 through 6, on the basis of multiple sequence alignment and phylogenetic tree reconstruction. Each family tends to have a somewhat similar ligand binding domain, but this varies widely between families (Mazaira et al., 2018; Weikum et al., 2018). Indeed, the ligand binding domains between families vary in terms of volume - some have a large pocket, others have a smaller, flexible, expanding pocket, and others still have a pocket already filled with side chains – as well as in terms of specificity, with some binding only specific ligands, others binding a range of ligands, and orphans where it is yet unknown what they bind, if anything (de Vera, 2018; de Vera et al., 2019; Markov & Laudet, 2011; Munoz-Tello et al., 2020). While the conservation across all families suggests that NRs share a common ancestor, debate remains over whether that ancestor was a ligand-activated NR or more like the orphan extant NRs, which may not be regulated by ligands (Markov & Laudet, 2011; Papageorgiou et al., 2021).

While the two more classical categories of NRs – binding a specific ligand, or having no known ligand – are what is generally discussed, two of the subfamilies have been found to exist outside of this dichotomy (Markov & Laudet, 2011). These are the subfamilies NR1, which tends to have a binding affinity for a variety of ligands, and NR4, which appears to have its ligand binding pocket filled by hydrophobic side-chains, leaving little to no room for ligands (Markov & Laudet, 2011; Munoz-Tello et al., 2020) – although one member of NR4 appears to have almost no binding pocket at all (Mazaira et al., 2018). Hydrophobic residues do often account for a large percentage of the ligand binding domain pocket, but it appears that only in the NR4 subfamily do they fill the pocket (Weikum et al., 2018). The NR1 and NR4 subfamilies, while appearing to function very differently to other NR subfamilies as well as to each other, appear to be more closely linked phylogenetically to each other than to the other subfamilies (Cheng et al., 2021; Papageorgiou et al., 2021). This suggests some similarity between them which separates them from the other five subfamilies, alongside some key difference which sets them apart from each other and provides them with such differing functions and almost opposing affinity for ligand binding.

Ancestral sequence reconstruction is a tool which allows study of the evolutionary history of protein families through the use of modern day (extant) organism protein sequences (Livada et al., 2023; Thomas et al., 2019). While the true ancestor sequence cannot be verified, as it likely no longer exists, ancestral sequences reconstruction allows inference of likely possibilities through generation of new sequences based upon probabilistic

searches of the non-conserved positions of a sequence. This, assuming the extant sequence alignment was accurate and relatively free of mutations, gives each output a high likelihood of being a functional sequence as the conserved residues are unlikely to change (Thomas et al., 2019). While reconstructed ancestor sequences can often vary from extants by up to thirty percent, the coordinated sets of mutations which can be found can allow identification of modified traits – most notably, increased stability under thermal or other stresses (Thomas et al., 2019). In a recent comparison of 113 protein sequences of which 56 were reconstructed ancestors, the reconstructions were found to be an average of 9ºC more stable than their extant counterparts, as well as to have equal or better thermostability than the closest extant neighbour in 94% of cases (Livada et al., 2023). Given the debate surrounding the ancestral state of NRs, the sequence similarity yet opposing ligand affinity of the NR1 and NR4 subfamilies, and the difficulty in determining NR structure due to instability, ancestral sequence reconstruction seems an apt choice for investigating NRs, by identifying positions of interest and potentially synthesising ancestors for structural investigation. This information in turn may help to open the way for improved therapeutic drugs.

However, while ancestral sequence reconstruction appears to be the solution, care must be taken with which ancestors are selected for use. Structural experiments may benefit from increased stability, but this can be achieved by even single amino acid substitutions. Assessing the effects of these mutations experimentally has the potential to be time consuming, laborious, and expensive (Blaabjerg et al., 2023). Instead, computational models have been developed which can predict not only protein stability, but also function, as these predictions continue to be an ongoing challenge (Blaabjerg et al., 2023; Cheng et al., 2024; Mansoor et al., 2023; Meier et al., 2021).

Two approaches are commonly taken in the attempt to predict stability: supervised and self-supervised models. In supervised models, predictions are made based on targets taken from experimental measures. While this means they can make predictions at a correct absolute scale, they tend to be susceptible to systematic bias through overfitting and bias towards experimentally induced mutations (Blaabjerg et al., 2023). In contrast to this, self-supervised models are trained using structure or sequence information to predict masked amino acids, and hence learning a likelihood distribution of amino acid types at various positions. This likelihood can then be used to predict the effects of mutations on stability (Blaabjerg et al., 2023). One example of this includes the method DeepSequence (Riesselman et al., 2018), which obtains high accuracy in mutation effect prediction after

training on multiple sequence alignments. However, other methods such as ProMEP and zero-shot predictors also perform well, with the advantage of not requiring specific training on a protein family of interest, but often at the cost of substantial computational resources and predictions on a non-absolute scale, which can limit practical applicability (Blaabjerg et al., 2023; Cheng et al., 2024; Mansoor et al., 2023). Methods which combine the two approaches have also been developed in order to select from the complementary strengths and weaknesses, with RaSP notable among them for its ability to rapidly calculate stability changes for a huge number of potential changes (Blaabjerg et al., 2023). One of the greatest difficulties in selecting the potential sequence positions to mutate for optimal results is the combinatorial explosion present, given because each position must be considered in conjunction with other positions due to their potential proximity in three-dimensional space. If methods are available to potentially predict the effect of a large number of these mutations, selecting sequences for further analysis becomes a more manageable task.

Project outline

The main aim of this project is to attempt to identify what ancestral mutations may have led the NR1 subfamily to evolve in a way that allowed for expansion of their ligand binding pockets, thus allowing them to accommodate a wide variety of ligands, while the NR4 subfamily evolved with a much smaller, hydrophobically packed pocket which discourages diverse ligand binding, while conserving the typical structure. This is of further interest given the sequence similarities between the two subfamilies which allow them to be classified so close together on a phylogenetic tree.

The main process which will be used to investigate these differing properties is ancestral sequence reconstruction. This will target three specific ancestors – the ancestor of the NR1 subfamily, the ancestor of the NR4 subfamily, and their joint ancestor. Candidates for each of these three ancestors will also be subject to molecular dynamic simulations, and potentially to investigations into their structures.
Because of the time and resources required to perform simulations or structural experiments, the choice of ancestor candidates must be carefully considered, especially in terms of identifying those which will have a higher potential for stability.
One way of identifying this may be to use existing stability predictor tools.

As part of the aim is to identify specifically which mutations may have led to the evolution of the two subfamilies, another line of computational investigation is to determine the effect of mutations at a variety of positions. Initial exploratory analyses show that extant sequences have their function correctly predicted by some tools. As such, it follows that certain aspects of the sequences are able to identify the sequences, and mutating these should disrupt that prediction. An investigation will be undertaken using existing tools to identify the predicted effect of mutating some positions both individually and as part of a set. Some key points it is hoped to identify through this are how many mutations are needed to change the predicted function, the relative importance of each site, and how can this information be used to take the potential combinatorial explosion of possible mutations down to a level which can be meaningfully interpreted by a predictor.

References

Blaabjerg, L. M., Kassem, M. M., Good, L. L., Jonsson, N., Cagiada, M., Johansson, K. E., Boomsma, W., Stein, A., & Lindorff-Larsen, K. (2023). Rapid protein stability prediction using deep learning representations. *Elife*, *12*. https://doi.org/10.7554/eLife.82593

Cheng, P., Mao, C., Tang, J., Yang, S., Cheng, Y., Wang, W., Gu, Q., Han, W., Chen, H., Li, S., Chen, Y., Zhou, J., Li, W., Pan, A., Zhao, S., Huang, X., Zhu, S., Zhang, J., Shu, W., & Wang, S. (2024). Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell Res*. https://doi.org/10.1038/s41422-024-00989-2

Cheng, Y., Chen, J., Mukhtar, I., & Chen, J. (2021). Genome-Wide Characterization of the Nuclear Receptor Gene Family in Macrostomum lignano Imply Its Evolutionary Diversification [Original Research]. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.653447

D'Arrigo, G., Autiero, I., Gianquinto, E., Siragusa, L., Baroni, M., Cruciani, G., & Spyrakis, F. (2022). Exploring Ligand Binding Domain Dynamics in the NRs Superfamily. *Int J Mol Sci*, *23*(15). https://doi.org/10.3390/ijms23158732

de Vera, I. M. S. (2018). Advances in Orphan Nuclear Receptor Pharmacology: A New Era in Drug Discovery. *ACS Pharmacol Transl Sci*, *1*(2), 134-137. https://doi.org/10.1021/acsptsci.8b00029

de Vera, I. M. S., Munoz-Tello, P., Zheng, J., Dharmarajan, V., Marciano, D. P., Matta-Camacho, E., Giri, P. K., Shang, J., Hughes, T. S., Rance, M., Griffin, P. R., & Kojetin, D. J. (2019). Defining a Canonical Ligand-Binding Pocket in the Orphan Nuclear Receptor Nurr1. *Structure*, *27*(1), 66-77 e65. https://doi.org/10.1016/j.str.2018.10.002

Livada, J., Vargas, A. M., Martinez, C. A., & Lewis, R. D. (2023). Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. *ACS Catalysis*, *13*(4), 2576–2585. https://doi.org/https://doi.org/10.1021/acscatal.2c03859

Mansoor, S., Baek, M., Juergens, D., Watson, J. L., & Baker, D. (2023). Zero-shot mutation effect prediction on protein stability and function using RoseTTAFold. *Protein Sci*, *32*(11), e4780. https://doi.org/10.1002/pro.4780

Markov, G. V., & Laudet, V. (2011). Origin and evolution of the ligand-binding ability of nuclear receptors. *Mol Cell Endocrinol*, *334*(1-2), 21-30. https://doi.org/10.1016/j.mce.2010.10.017

Mazaira, G. I., Zgajnar, N. R., Lotufo, C. M., Daneri-Becerra, C., Sivils, J. C., Soto, O. B., Cox, M. B., & Galigniana, M. D. (2018). The Nuclear Receptor Field: A Historical Overview and Future Challenges. *Nuclear receptor research*, *5*. https://doi.org/https://doi.org/10.11131/2018/101320

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. [https://papers.nips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf]. 35th Conference on Neural Information Processing Systems (NeurIPS), Online.

Munoz-Tello, P., Lin, H., Khan, P., de Vera, I. M. S., Kamenecka, T. M., & Kojetin, D. J. (2020). Assessment of NR4A Ligands That Directly Bind and Modulate the Orphan Nuclear Receptor Nurr1. *J Med Chem*, *63*(24), 15639-15654. https://doi.org/10.1021/acs.jmedchem.0c00894

Papageorgiou, L., Shalzi, L., Pierouli, K., Papakonstantinou, E., Manias, S., Dragoumani, K., Nicolaides, N. C., Giannakakis, A., Bacopoulou, F., Chrousos, G. P., Eliopoulos, E., & Vlachakis, D. (2021). An updated evolutionary study of the nuclear receptor protein family. *World Academy of Sciences Journal*, *3*(6). https://doi.org/https://doi.org/10.3892/wasj.2021.122

Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*, *15*(10), 816-822. https://doi.org/10.1038/s41592-018-0138-4

Rodriguez-Calvo, R., Tajes, M., & Vazquez-Carrera, M. (2017). The NR4A subfamily of nuclear receptors: potential new therapeutic targets for the treatment of inflammatory diseases. *Expert Opin Ther Targets*, *21*(3), 291-304. https://doi.org/10.1080/14728222.2017.1279146

Shulman, A. I., Larson, C., Mangelsdorf, D. J., & Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell*, *116*(3), 417-429. https://doi.org/10.1016/s0092-8674(04)00119-9

Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M., & Harmer, N. (2019). Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol*, *2*, 429. https://doi.org/10.1038/s42003-019-0677-y

Weikum, E. R., Liu, X., & Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Sci*, *27*(11), 1876-1892. https://doi.org/10.1002/pro.3496