Domain expansion and functional diversification in vertebrate reproductive proteins

Alberto M. Rivera[1], Damien B. Wilburn[1,2], Willie J. Swanson[1]

1. Department of Genome Sciences, University of Washington
2. Department of Biomedical Informatics, The Ohio State University

Abstract

The rapid evolution of fertilization proteins can result in remarkable diversity in their structure and function. Many of the proteins in vertebrate egg coats contain copies (1-6) of the ZP-N domain. These ZP-N domains can facilitate multiple reproduction functions, including species-specific sperm recognition. We integrated phylogenetics and machine learning to investigate how ZP-N domains diversified in structure and function. The most C-terminal ZP-N domain of each paralog is associated with another domain type (ZP-C) which together form a "ZP module." All modular ZP-N domains were phylogenetically distinct from non-modular or free ZP-N domains. Machine learning-based classification identified 8 residues that form a stabilizing network in modular ZP-N domains that is absent in free domains. Positive selection was identified in some free ZP-N domains. Our findings suggest that purifying selection has conserved an essential structural core in modular ZP-N domains, while free N-terminal domains have been able to experience functionally diversify.

Introduction

One of the evolutionary innovations associated with multicellularity is the specialization and organization of cells into tissues and organs. Construction of such multicellular structures requires cells to manipulate their local environments by the secretion of different proteins. These components assemble into extracellular matrices and contribute to tissue functions such as morphogenesis, tissue repair (Clause and Barker 2013), homeostasis, and cell adhesion (Streuli 2016). Such extracellular proteins (e.g. collagen) often contain peptide motifs that allow assembly into fibrils and higher order structures (Brodsky and Persikov 2005). Extracellular matrices may be diversified by the duplication of whole genes and individual protein domains, providing new targets for natural selection to act upon. Vertebrate egg coat proteins present a clear example where a complex history of domain duplication is tied to the evolution of an essential extracellular structure.

Vertebrate oocytes are surrounded by an elevated glycoprotein envelope that provides protection from the environment and regulates the rate of sperm entry (Monne et al. 2008). This structure goes by many names, termed the zona pellucida (ZP) in mammals, the chorion in fishes, and the vitelline membrane in amphibians, reptiles, and birds (Wilburn and Swanson 2018). Named after the mammalian version, all vertebrate egg coat proteins contain a pair of immunoglobulin-like domains, ZP-N and ZP-C, that together form a polymerization unit called a ZP module (Jovine et al. 2002; Wilburn and Swanson 2017; Bokhove and Jovine 2018). The last common ancestor of vertebrates possessed six paralogous genes (*zp1*, *zp2*, *zp3*, *zp4*, *zpd*, *zpax*) that have experienced clade-specific birth and death events. Consequently, the egg coat of each major vertebrate class has a different composition of ZP module-containing proteins (Conner et al. 2005; Wong and Wessel 2005; Goudet et al. 2008; Meslin et al. 2012; Shu et al. 2015; Wassarman and Litscher 2016; Killingbeck and Swanson 2018). ZP modules are also found in non-reproductive proteins that form extracellular matrices, such as uromodulin (UMOD) which protects against urinary pathogens (Brunati et al. 2015; Bokhove et al. 2016; Devuyst and Pattaro 2018) and tectorin alpha (TECTA) which function in inner ear organization (Bokhove et al. 2016; Kim et al. 2019).

While both ZP-N and ZP-C are immunoglobulin-like domains with a core β-sandwich (Bokhove and Jovine 2018), they are evolutionarily distinct domains that have low amino acid

sequence identity, unique disulfide patterns, and variable loop structures (Lin et al. 2011). Independent ZP-C domains outside of the ZP module have been identified in *C. elegans* (Weadick 2020), and four of the egg coat proteins (ZP1, ZP2, ZP4, and ZPAX) contain additional ZP-N domains independent of the ZP-N/ZP-C pair in the ZP module. We refer to ZP-N domains in the module as “modular” and the N-terminal repeats as “free” domains. As ZP-N domains can form asymmetric dimers through their β-sandwich edges (Jovine et al. 2002; Bokhove and Jovine 2018; Litscher and Wassarman 2020), they have been considered the major driver of ZP module polymerization. While free ZP-N domains may similarly function as polymerization units, recent structural studies support that they may have acquired novel functions: the free ZP-N domains of ZP1 form intermolecular cross-links important for egg coat structure (Nishimura et al. 2019), while N-terminal domains in ZP2 (Avella et al. 2013; Avella et al. 2014) and ZP4 (Dilimulati et al. 2022) have been implicated in sperm-egg binding. Despite their functional significance, the evolutionary history of ZP-N domains within and between these many paralogous proteins has not been examined. Using a combination of phylogenetic and machine learning approaches, this manuscript addresses how a complex history of whole gene and tandem domain duplications followed by structural adaptation produced the current diversity of ZP proteins.

Results and Discussion

We investigated the evolutionary history of vertebrate ZP-N domains by extracting a total of 1448 ZP-N domain sequences from ZP module containing genes of 210 species with both reproductive (*zp1*, *zp2*, *zp3*, *zp4*, *zpax*, *zpd*) and non-reproductive (*umod*, *tecta*, *cuzd1*) functions (Table S1). While modular and free ZP-N sequences were found to share little sequence identity beyond 4 conserved cysteine residues that form stabilizing disulfide bonds, both domain types are highly similar in three-dimensional structure (Fig 1A). As such, we used a structure-based sequence alignment (Pei et al. 2008) to perform phylogenetic analysis. Maximum likelihood-based phylogenies indicated that the free ZP-N domains form a single clade distinct from the ZP-C associated modular ZP-N domains (Fig 1B), and this separation was robust to amino acid substitution matrices (LG, WAG, and JTT) (Fig S1). The topology of the modular ZP-N clade was broadly consistent with previously published gene trees based on the complete ZP module with both ZP-N and ZP-C (Claw and Swanson 2012; Feng et al. 2018). The topology of the free ZP-N clade supports that the initial duplication gave rise to the first repeat of the tandem array shared by ZP1, ZP2, ZP4, and ZPAX, which was followed by lineage specific repeat expansions of free ZP-Ns in *ZP2* and *ZPAX* (Fig 1C).

The phylogenetic separation of modular and free ZP-N domains using a structure-based alignment suggests important structural differences between the two domain types, but their high sequence divergence complicated manual identification of such characteristics. Machine learning methods have been applied to various aspects of protein biology such as function prediction (Yang et al. 2018; Bonetta and Valentino 2020) and the classification of membrane bound proteins (Guo et al. 2019). Here, we used a machine learning-based classification strategy to identify what structural features distinguish these two types of ZP-N domains. We applied a logistic regression model to the structurally aligned ZP-N domain sequences, where the probability of being a modular vs free ZP-N type was estimated for each of the 20 amino acids at each position in the alignment. Given the large number of parameters in this model (7981), we combined elastic net regularization and cross-validation to identify the most

parsimonious model (i.e. the fewest non-zero parameters) within the 95% confidence interval of the highest scoring model (Fig 2A-B). Through this regularization strategy, we identified a total of 8 modular-associated residues that were sufficient to predict whether a given ZP-N sequence was modular or free with 100% accuracy (Fig 2B).

Examination of the residues associated with either ZP-N type in the context of three-dimensional structures suggest differences in both function and quaternary structural dynamics. ZP-N monomers have an immunoglobulin-like β-sandwich fold with the 4- and 3-membered β-strands connected by a disulfide bridge on each edge of the molecule. Biochemical and crystallographic studies support that modular ZP-N domains form asymmetric dimers through the molecular edge that includes the most N- and C-terminal β-strands (Jovine, et al. 2006; Bokhove, et al. 2016). Free ZP-N domains do not appear to dimerize through this N/C-terminal edge, and have experienced functional diversification of the outer edge of the molecule to perform additional protein binding functions (Raj, et al. 2017; Nishimura, et al. 2019). When the modular-associated sites were mapped onto their respective structures, we observed that modular-associated residues form an integrated network of mostly hydrophobic stabilizing contacts that interlock between the β-sheets around the outer edge of the molecules (Fig 2C, Fig S2). The phylogenetic clustering of free ZP-N domains (Fig 1C), along with molecular dynamics support the loss of dimerization activity along the free ZP-N lineage, which could have facilitated their evolution of new binding partners (Fig S3). The stabilizing contacts along the outer edge of the modular ZP-N domains are consistent with these domains principally having structural roles, while in free domains this edge has diversified to allow functional innovation. Further subdivision of free ZP-N domains by their major clades (the first repeat versus internal repeats in ZP2 and ZPAX) provided little additional information distinguishing these different free ZP-Ns from one another or modular ZP-N domains (Fig S4). Consequently, our sequence-based machine learning classifier appears to have identified residues underlying structural differences between the two domain types that have implications on their respective functions.

The difference in relative conservation of modular domain structures motivated additional analysis of the sequence evolution of these ZP-N domains. Here we focused on mammalian ZP genes (*zp1*, *zp2*, *zp3* ,*zp4*, *umod*, *tecta*, and *cuzd1*) due to both higher genomic assembly quality and to avoid synonymous substitution saturation that may occur when considering greater phylogenetic breadth (Anisimova and Liberles 2012). Measures of sequence diversity within and between ZP-N groups reveal that modular domains are less diverse overall, and that free ZP-Ns are just as dissimilar to one another as they are to modular domains (Fig 3A).

These findings motivated molecular evolutionary analyses, and of the 12 ZP-N domains analyzed, only ZP2-N1 and ZP2-N2 showed evidence of positive selection (Table S2). These are notably the two domains with the lowest within group similarity (diagonal of Fig 3A). Positively selected sites in ZP2-N1 are far from the homodimerization edge and physically closer to the network of modular biased residues (Fig 3B). Positively selected sites also constituted a substantial portion of the solvent exposed surface area (34% in ZP2-N1 and 24% in ZP2-N2), potentially facilitating their evolution of novel functions and protein interactions. The rapid evolution of ZP2-N1 is consistent with its role in species-specific sperm recognition (Avella, et al. 2014) and may reflect sexual coevolution with its sperm receptor (whose identity is currently unknown). Remarkably, these positively selected sites cluster near a region associated with species-specific sperm protein binding in free invertebrate ZP-N domains (Raj et al. 2017).

However, based on expansion and retraction of loop lengths outside the core β-sandwich, we believe that these invertebrate free ZP-N domains evolved independently of the free ZP-N domains of vertebrates, suggesting that the expansion of ZP-N arrays for species-specific sperm recognition is a convergent phenomenon that has arisen multiple times throughout metazoan evolution.

In summary, our combined phylogenetic, machine learning classification, and positive selection analyses illustrated a clear distinction between modular and free ZP-N domains. These two classes of domains experienced different evolutionary trajectories, as modular ZP-Ns likely retained a conserved structural role while free ZP-Ns neofunctionalized to serve different reproductive functions. These findings are of relevance to the evolution of species-specificity in fertilization, as the ZP-N domain expansion of ZP2 provided substrates to evolve novel species-specific interactions. Structural changes within free ZP-Ns could result in of a dimerization edge and the evolution of a new sperm binding loop. As these domains are coopted into a reproductive context, coevolution (Clark et al. 2009; Hart et al. 2018) and sexual conflict (Gavrilets and Waxman 2002) with sperm proteins could contribute to their rapid evolution. This reflects the evolutionary dynamics that drive structural diversification and neofunctionalization of duplicated domains. Our combined phylogenetic and machine learning approach outlined here can be applied to other essential gene families with complex duplication histories.

Materials and Methods

Multiple Sequence Alignment

Sequences for multiple ZP-N containing proteins were curated from the Ensembl database (release 104) (Howe et al. 2021). Sequences were assigned labelled as one of the ZP genes of interest based on PSI-BLAST e-value scores (Altschul et al. 1997). Sets of orthologous genes were aligned preliminarily with MAFFT (Katoh and Standley 2013) and then trimmed to individual ZP-N domains. Groups of orthologous ZP-N domains were deemed “orthogroups”. Sequences with ambiguous characters were removed, and then sets of orthologous ZP-N sequences were realigned with MAFFT. A full multiple sequence alignment was made by putting the individual orthogroup alignments together using a representative paralog alignment. Individual representative sequences were taken from each orthgroup, and these paralogs were aligned using the structural based PROMALS tool (Pei et al. 2008). This approach was used because of the low sequence identity, but high structural similarity between paralogous Z-N domains. A custom script was used to algorithmically add gaps to orthogroup alignments to form a full multiple sequence alignment. CD-Hit was used to remove highly cluster highly similar sequences (>90% identity) (Li and Godzik 2006; Fu et al. 2012), in order to improve computing speed, and because this study was not concerned with very recent phylogenetic splits.

Phylogenetics

Maximum likelihood phylogenies were built using RAxML-NG(Kozlov et al. 2019), and multiple different amino acid substitution matrices were tested (LG+G, JTT+G,WAG+G), to demonstrate the robusticity of the deepest phylogenetic divide. The best tree was selected from 100 replications of the maximum likelihood analyses. Nodal support was calculated with transfer bootstrap expectation (Lemoine et al. 2018), a modified form of bootstrapping that is more effective at detecting deep phylogenetic relationships in datasets with large number of taxa. While initial labelling of sequences was based on BLAST values, but when these labels were ambiguous labelling was based on phylogenetic clustering. The clades of ZP1-N1, ZP2-N1, and ZP4-N1, were labelled according to a 90% ma

Machine Learning

A basic machine learning algorithm using mean squared regression and regularization was coded on python to distinguish the two free and modular groups of ZP-N domains. Logistic regression models are well suited for these classifications, because their outputs are bounded between 0 and 1, which can be interpreted as probability that a given domain is modular(Bewick et al. 2005). The multiple sequence alignment was identical to that used for phylogenetic analysis. The alignment was split into a testing (25%) and training set (75%), and we employed logistic regression modelling with cross-validation on the training set. In that approach, the training set was split into five subsets, and then each of the five subsets is used to validate models trained on the other rest of the training data. For additional rigor, the separate testing dataset was used for the final scoring and model selection.

In order to encode an aligned ZP-N domain within this machine learning framework, each position in the sequence was converted into a vector of twenty digits, corresponding to the twenty amino acids. The value was set to 1 for the entry in the vector corresponding to that residue, and all other values are set to 0. Gapped sites were set to a vector of twenty 0’s. Thus, the classifier was trained using 20n features (plus an additional intercept term), where n is the

alignment length. Each of these features has a parameter associated with it and the value of the parameter indicates how informative that feature is, and whether it supports a modular ZP-N or free ZP-N classification. There are a large number of possible parameters in this model (7981 including the intercept), but it is worth considering the optimal number of non-zero parameters for this model. Increasing the number of non-zero parameters will always improve its accuracy on the data it was trained on, but too many parameters will reduce its accuracy when applied to other datasets, in a phenomenon called “overfitting.”(Hawkins 2004)

In optimizing these models, we employed elastic net regularization which two functions that penalize parameters and reduce risk of overfitting (Zou and Hastie 2005). In our sci-kit learn implementation (Pedregosa et al. 2011), we varied both the strength of regularization and the ratio between the two penalty types, and evaluated a range of possible models. The highest scoring model was identified according to the negative mean-squared error scoring metric. In order to choose a suitable sparse model (i.e. fewest non-zero parameters), we adapted the one standard error rule common in machine learning (Hastie et al. 2009), in which you select the sparsest model that is still within one standard error of the highest scoring model. For this analysis we used 95% confidence intervals (~1.96 standard errors), to achieve the sparsest model (fewest non-zero parameters) that is not statistically different from the highest scoring model sampled. Raw parameter values were plotted in the style of sequence LOGO plots (Schneider and Stephens 1990). The sum of the raw parameter values for matching amino acids in the alignment (and the intercept term), are essentially the log odds that a given sequence is classified as modular. For simplicity, each parameter is described as the log odds associated with a particular residue. After the phylogenetic analysis separated the first N-terminal ZP-N as its own clade, we re-ran the analysis with three multiclassification, but with an otherwise identical hyperparameter search space and computational pipeline.

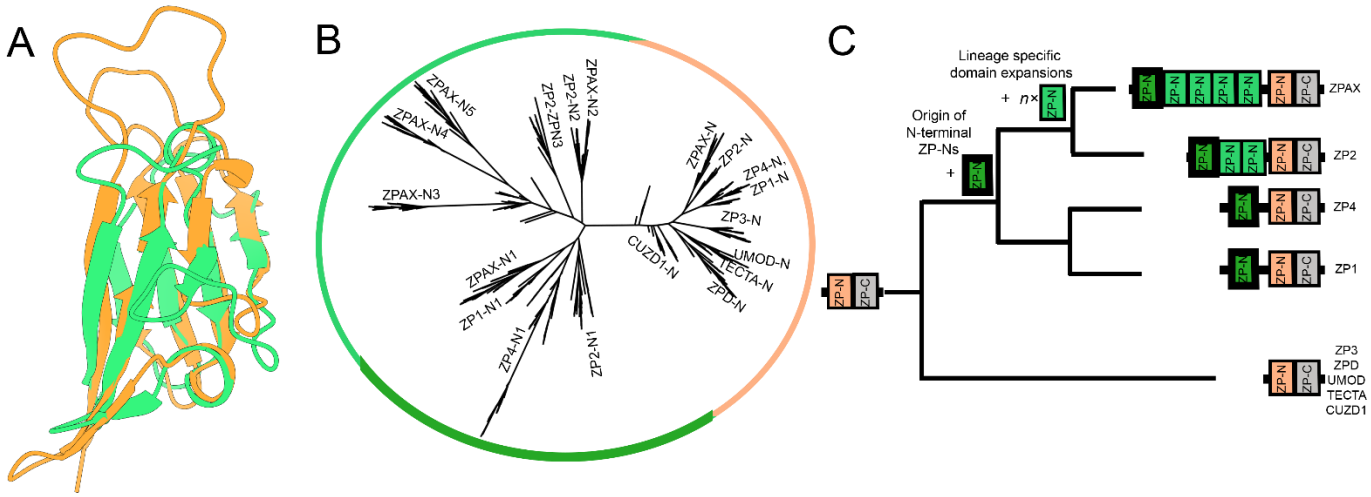
Sequence Divergence and Positive Selection Analyses

Our analyses of sequence divergence and positive selection was limited to boreutherian mammals, because of genome quality concerns, and because at large phylogenetic distances sequence differences begin to approach saturation which obscures evolutionary differences (Anisimova and Liberles 2012). For this reason, we were limited to the 12 mammalian ZP-N domains coming from *zp1*, *zp2*, *zp3*, *zp4*, *umod*, *tecta*, and *cuzd1*. Boreoeutherian sequences were mined from ensemble (Howe et al. 2021), and were included in these analyses if they were present in 10 or more of these ZP-N domain orthogroups. Phylogenetic distances both within and between orthogroups were calculated in MEGA using poisson estimation with a gamma distribution of variation between sites (Kumar et al. 2016; Kumar et al. 2018).

To determine whether there was any evidence of positive selection we ran PAML analyses (Yang et al. 2005; Yang 2007) on the same sets of ZP-N domains from the sequence divergence estimation. A likelihood ratio test between a model allowing positive selection (M8) and a neutral model (M8a), was used to determine which domains showed evidence of positive selection. Twice the difference of log-likelihood between the models was applied to a chi-squared distribution with one degree of freedom, because M8 has one more parameter M8a. We also performed a Benjamini-Hochberg p-value correction to account for multiple testing (Benjamini and Hochberg 1995). Positively selected sites were visualized on a published crystal structure (ZP2-N1) (Raj et al. 2017), or the alpha-fold predicted structure (Jumper et al. 2021 Jul 15) when this did not exist (ZP2-N2). Sites were labelled if they had a posterior probability of being positively selected > 75% according Bayes Empirical Bayesian (BEB) analysis.

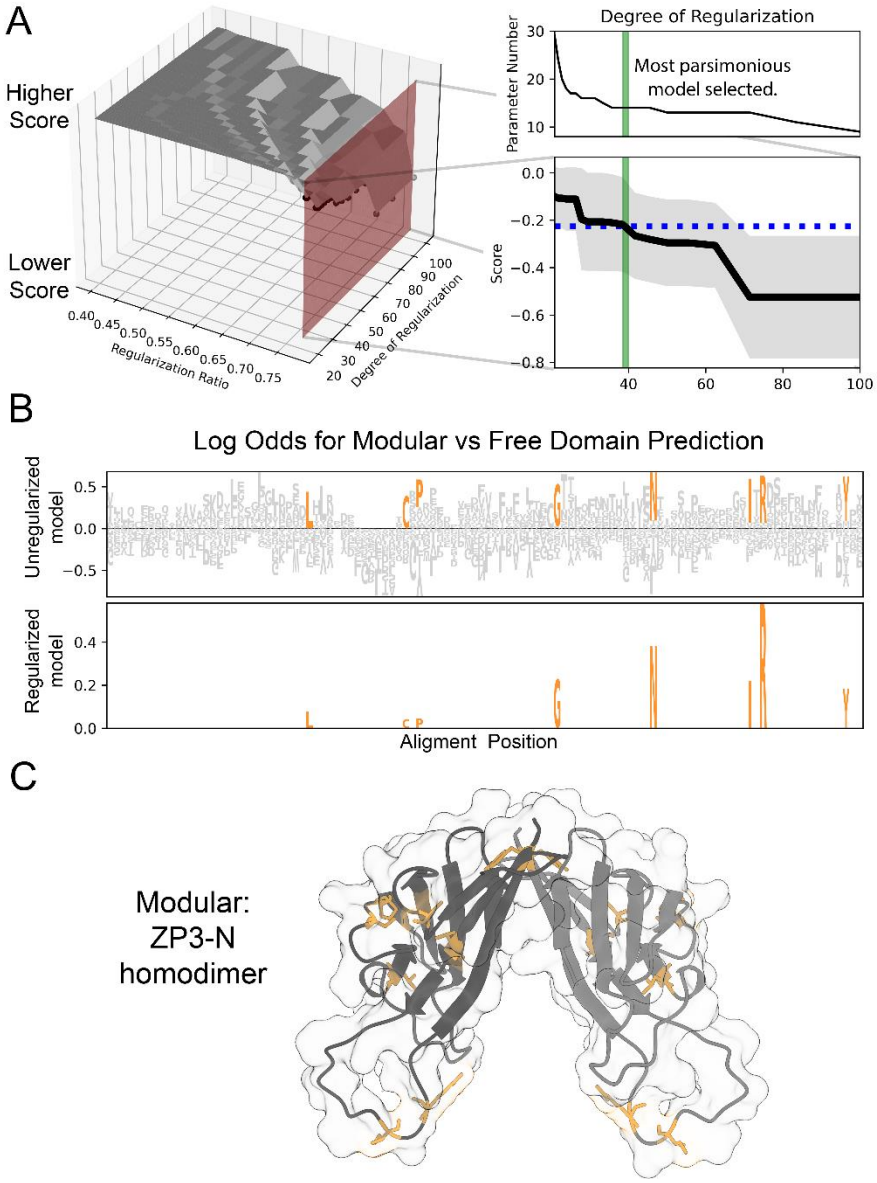
Visualization and Other methods

When protein structures were not available Alpha-Fold2 tertiary structure prediction was used (Jumper et al. 2021), and three-dimensional protein structures were visualized using either pymol (Schrödinger 2015) or ChimeraX (Pettersen et al. 2004). Rosetta (Chaudhury and Gray 2008; Sircar et al. 2010) was used to perform docking simulations on ZP2-N1 and ZP3-N homodimers, to illustrate the greater energetic stability of modular ZP-N homodimers.



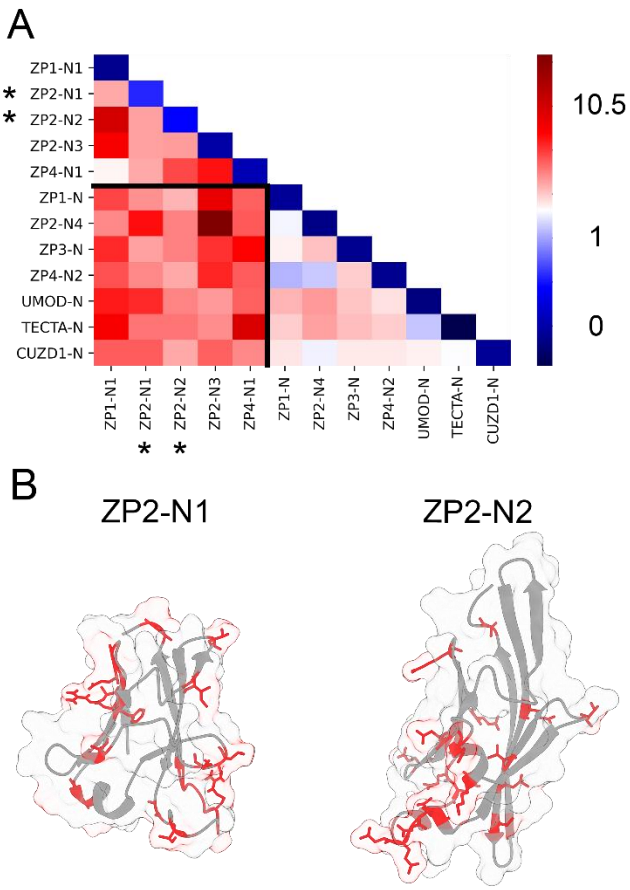
Phylogenetic analysis of ZP-N domain duplication history

Figure 1: A) A structural alignment of Mouse ZP2-N1 and ZP3-N highlights the broad structural conservation of these two classes of ZP-N domains (RMSD = ~4.7 Å) despite only ~18% amino acid sequence identity. B) Phylogenetic analysis (Kozlov et al. 2019) of ZP-N sequences (shown as a maximum likelihood tree) supports an ancestral separation between free and modular ZP-N domains (~78% support). C) A summary of ZP-N domain evolution based on the gene tree in panel B. The ancestral protein contained a ZP module with a C-terminal ZP-N and ZP-C domains, and duplication of the ZP-N produced the most N-terminal domain found in ZP1, ZP4, ZP2, and ZPAX. Later duplication events within ZP2 and ZPAX gave rise to multiple additional ZP-N domains between ZP-N1 and the ZP module.



Machine learning based inference of sequence features that distinguish modular and free ZP-N domains.

Figure 2: A logistic regression model with elastic net regularization was trained on the ZP-N multiple sequence alignment generated as part of the phylogenetic analysis, with the data partitioned for training and testing (75% and 25%, respectively), with 5-way cross validation of the training data employed to estimate the error distribution of the score function. We defined our optimal model as the most parsimonious model (i.e. the fewest parameters) within the estimated 95% confidence interval of the unregularized model. (A) The space of regularization hyperparameters was explored during model optimization, plotted as a 3D surface (left). The score is the negative mean squared error, and dots correspond to the two-dimensional cross section shown on the right, with the blue line denoting the intersection between the lower confidence limit of the unregularized model to its intersection with the score as a function of regularization strength. B) Comparison of the unregularized and optimal logistic regression models as LOGO plots with the height of each amino acid at each position corresponding to its parameter weight, with colored amino acids denoting parameters retained in the regularized model. Each parameter weight approximating the logs odd ratio for a modular domain prediction, when a residue is present at that position. C) Mapping highly predictive sites onto ZP-N protein models suggest differences in structural properties between free and modular domain. The available crystal structure ZP3-N (3d4c) was used and modelled as a dimer for spatial context. Modular-associated sites are generally buried along the outer edge of the homodimer.



Amino acid diversity and tests of positive selection in modular and free ZP-N domains.

Figure 3: A) A heatmap showing the within group and between group mean phylogenetic distances for orthologous groups of ZP-N domains (Kumar et al. 2018). B) Positively selected sites in mammalian ZP2-N1 and ZP2-N2 were identified through maximum likelihood analysis and mapped onto protein models (4wrn for ZP2-N1, and an AlphaFold prediction for ZP2-N2) (Yang 2007).

	Mammals	Birds	Reptiles	Amphibians	Fish
ZP1-N1	30	11	10	1	19
ZP1-N	30	9	8	0	20
ZP2-N1	46	20	4	2	34
ZP2-N2	46	19	10	1	21
ZP2-N3	45	18	11	2	3
ZP2-N	38	7	10	1	14
ZP3-N	36	13	12	2	38
ZP4-N1	37	11	10	2	23
ZP4-N	29	8	8	1	42
ZPAX-N1	2	9	11	2	50
ZPAX-N2	2	9	9	2	53
ZPAX-N3	2	9	9	2	51
ZPAX-N4	2	8	7	2	56
ZPAX-N5	2	12	11	2	55
ZPAX-N	2	5	8	2	44
ZPD-N	0	9	10	2	54
UMOD-N	41	0	6	1	6
TECTA-N	1	1	0	0	2
CUZD1-N	38	23	13	2	37

Summary of species sampled in phylogeny

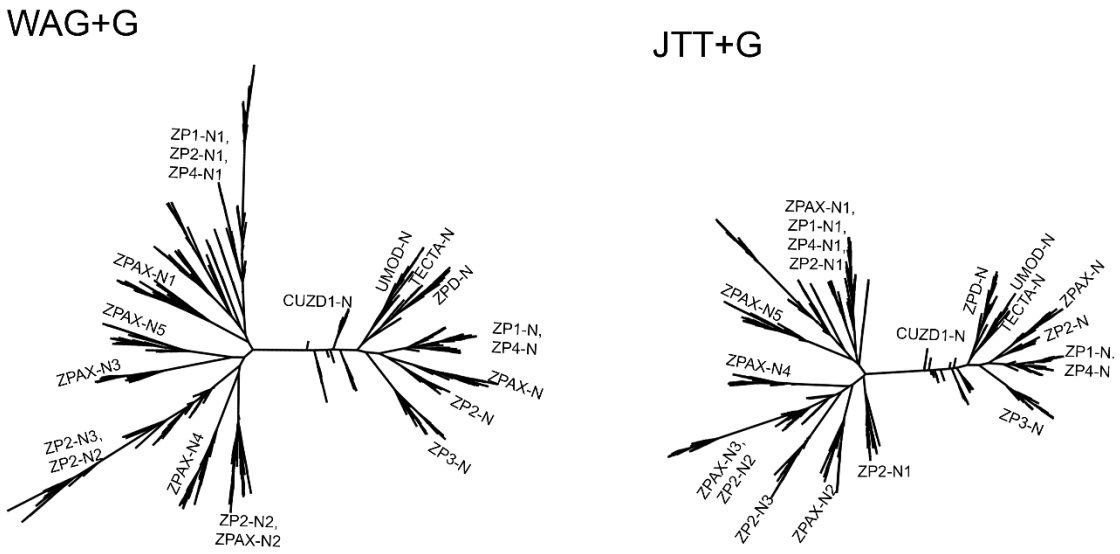
Table S1: We included a total of 1488 ZP-N domain sequences across the five classes of vertebrates in the final phylogeny. Here the amphibians only included frogs due to genomic availability reasons. The fish class includes all non-tetrapod vertebrates, and is non-monophyletic. Labels in this table are based on BLAST results.

Domain	M8a	M8	-2ΔlogL	*p* value	*p* value corrected
CUZD1-N	p_0 = 0.97, p_1= 0.03, p = 0.64, q=2.1, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.63, q=1.9, ω = 2.5	2.4	0.06	0.25
TECTA-N	p_0 = 0.98, p_1= 0.02, p = 0.05, q=0.99, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.05, q=0.97, ω = 1.9	1.0	0.16	0.24
UMOD-N	p_0 = 0.98, p_1= 0.02, p = 0.29, q=0.61, ω = 1	p_0 = 0.98, p_1= 0.02, p = 0.29, q=0.61, ω = 1	0	0.5	0.57
ZP1-N1	p_0 = 0.74, p_1= 0.26, p = 0.74, q=1.4, ω = 1	p_0 = 0.99, p_1= 0.01, p = 0.54, q=0.51, ω = 3.5	1.9	0.08	0.2
ZP1-N2	p_0 = 1, p_1= 0, p = 0.46, q=0.68, ω = 1	p_0 = 1, p_1= 0, p = 0.46, q=0.68, ω = 1	0	0.5	0.57
ZP2-N1	p_0 = 0.43, p_1= 0.57, p = 2.7, q=8.0, ω = 1	p_0 = 0.65, p_1= 0.35, p = 1.2, q=1.4, ω = 1.5	6.7	4.7E-03	0.03 *
ZP2-N2	p_0 = 0.42, p_1= 0.58, p = 29.5, q=99, ω = 1	p_0 = 0.70, p_1= 0.30, p = 1.6, q=1.5, ω = 1.9	18.8	7.2E-06	8.6E-5 *
ZP2-N3	p_0 = 0.67, p_1= 0.33, p = 1.6, q=7.9, ω = 1	p_0 = 0.85, p_1= 0.15, p = 0.65, q=1.5, ω = 1.4	1.5	0.11	0.22
ZP2-N	p_0 = 0.87, p_1= 0.13, p = 0.76, q=3.4, ω = 1	p_0 = 0.87, p_1= 0.13, p = 0.76, q=3.4, ω = 1	0	0.5	0.57
ZP3-N	p_0 = 0.98, p_1= 0.02, p = 0.55, q=1.2, ω = 1	p_0 = 1, p_1= 0, p = 0.52, q=1.21 ω = 2.3	0	0.5	0.57
ZP4-N1	p_0 = 0.58, p_1= 0.42, p = 1.3, q=4.3, ω = 1	p_0 = 0.93, p_1= 0.07, p = 0.64, q=0.70, ω = 1.9	2.3	0.065	0.2
ZP4-N	p_0 = 0.82, p_1= 0.18, p = 0.40, q=1.1, ω = 1	p_0 = 0.94, p_1= 0.06, p = 0.37, q=0.67, ω = 1.5	1.1	0.15	0.26

Summary of PAML Output

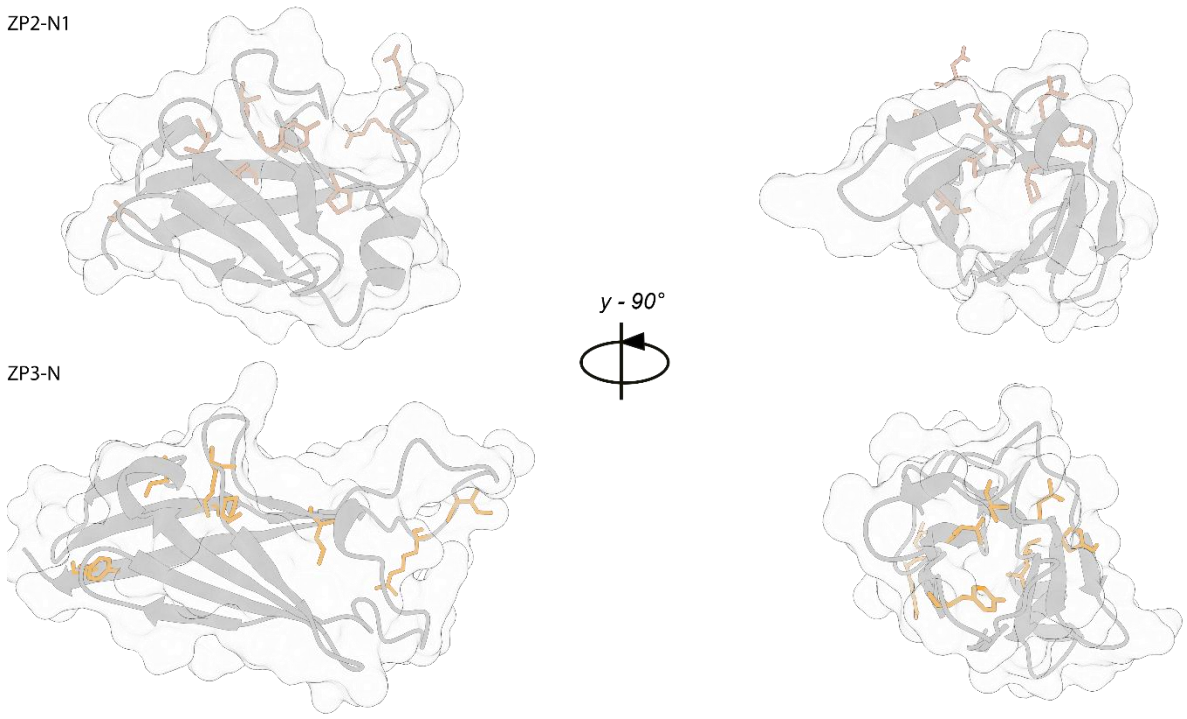
Table S2: This table summarizes the results from PAML analysis (Yang 2007). Here we compared a neutral model (M8a) to a model that allows positive selection (M8). A * denotes statistically significant p values after Benjamini-Hochberg multiple testing correction (Benjamini and Hochberg 1995).

\



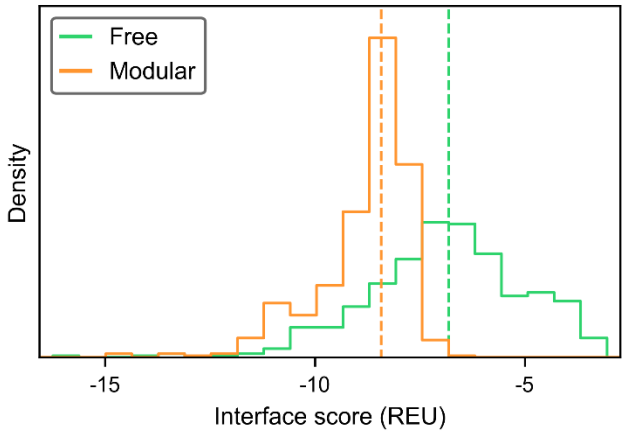
Phylogenies from alternative substitution matrices

Supplemental Figure 1: These two trees were produced through RAxML-NG(Kozlov et al. 2019). Major aspects of the tree are conserved, specifically the monophyletic grouping of free ZP-N domains.



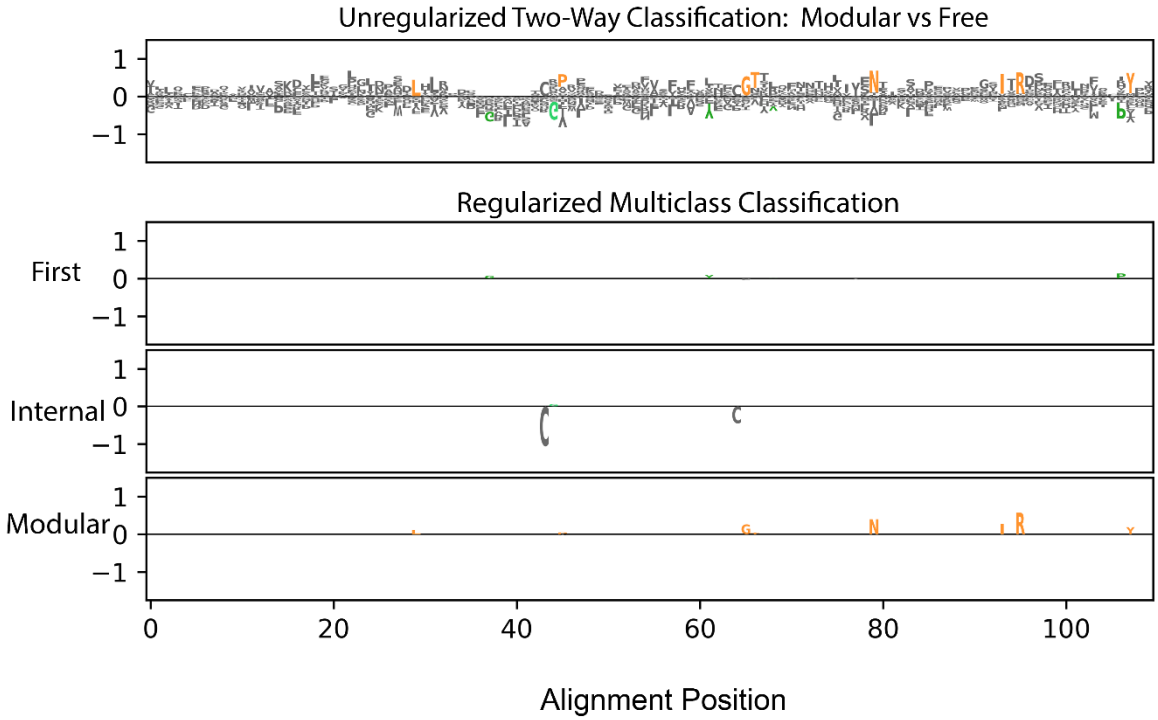
Modular-biased sites in the context of the ZP-N domain core

Supplemental Figure 2: This figure shows the modular biased sites in the context of both ZP2-N1 (Raj et al. 2017) and ZP3-N (Monne et al. 2008) crystal structures. The brighter orange corresponds to the modular-biased sites according to our machine learning model, while the duller orange on ZP2-N1 represent the structural homolog of those sites according to a pymol (Schrödinger 2015) structural alignment. We observe a clustering of these sites within the core of the domain in ZP3-N and not ZP2-N1. The two positions of each are 90° rotations around the y-axis.



Rosetta docking simulation of dimerization

Supplemental Figure 3: This is a histogram of interface scores from rosetta docking simulations of dimers for both free (ZP2-N1) and modular (ZP3-N) ZP-N domains.



Multiclass machine-learning ZP-N domain classification

Supplemental Figure 4: Since we observed monophyletic grouping of the most N-terminal ZP-N domain, we performed a multiclass variation of our machine learning analysis. The data was split into three classes: modular, first (i.e. most N-terminal free domain), and internal domains. In this multiclass analysis the model is fit three times, each time producing a classifier that distinguishes one of the classes from the rest of the data. The first row is the unregularized model from our two-class analysis, and our regularized multiclass models are summarized in the other three rows, where positive values suggest a bias towards that class. Our modular classifier recapitulates our earlier results, because it is in essence still comparing modular ZP-Ns versus all free ZP-Ns. The first ZP-N domain has few amino acids associated with it with relatively low parameter values. The internal ZP-N domains seem to have a bias against the cys 2-3 bond, but that likely is a reflection of the conservation of the second cysteine in the first ZP-Ns and the third cysteine in modular domains

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25(17):3389–3402. doi:10.1093/nar/25.17.3389.

Anisimova M, Liberles D. 2012. Detecting and understanding natural selection.

Avella MA, Baibakov B, Dean J. 2014. A single domain of the ZP2 zona pellucida protein mediates gamete recognition in mice and humans. J Cell Biol. 205(6):801–809. doi:10.1083/jcb.201404025.

Avella MA, Xiong B, Dean J. 2013. The molecular basis of gamete recognition in mice and humans. Mol Human Reprod. 19(5):279–289.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Bewick V, Cheek L, Ball J. 2005. Statistics review 14: Logistic regression. Crit Care. 9(1):112–118. doi:10.1186/cc3045.

Bokhove M, Jovine L. 2018. Structure of Zona Pellucida Module Proteins. Current Topic in Developmental Biology.(In Press).

Bokhove M, Nishimura K, Brunati M, Han L, de Sanctis D, Rampoldi L, Jovine L. 2016. A structured interdomain linker directs self-polymerization of human uromodulin. Proc Natl Acad Sci USA. 113(6):1552. doi:10.1073/pnas.1519803113.

Bonetta R, Valentino G. 2020. Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics. 88(3):397–413. doi:10.1002/prot.25832.

Brodsky B, Persikov AV. 2005. Molecular Structure of the Collagen Triple Helix. In: Advances in Protein Chemistry. Vol. 70. Academic Press. p. 301–339. https://www.sciencedirect.com/science/article/pii/S0065323305700097.

Brunati M, Perucca S, Han L, Cattaneo A, Consolato F, Andolfo A, Schaeffer C, Olinger E, Peng J, Santambrogio S, et al. 2015. The serine protease hepsin mediates urinary secretion and polymerisation of Zona Pellucida domain protein uromodulin. Elife. 4:e08887–e08887. doi:10.7554/eLife.08887.

Chaudhury S, Gray JJ. 2008. Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. Journal of Molecular Biology. 381(4):1068–1087. doi:10.1016/j.jmb.2008.05.042.

Clark N, Gasper J, Sekino M, Springer S, Aquadro C, Swanson W. 2009. Coevolution of Interacting Fertilization Proteins. PLoS Genet. 5(7):e1000570.

Clause KC, Barker TH. 2013. Extracellular matrix signaling in morphogenesis and repair. Current Opinion in Biotechnology. 24(5):830–833. doi:10.1016/j.copbio.2013.04.011.

Claw KG, Swanson WJ. 2012. Evolution of the Egg: New Findings and Challenges. Annu Rev Genom Hum Genet. 13(1):109–125. doi:10.1146/annurev-genom-090711-163745.

Conner SJ, Lefièvre L, Hughes DC, Barratt CLR. 2005. Cracking the egg: increased complexity in the zona pellucida. Human Reproduction. 20(5):1148–1152. doi:10.1093/humrep/deh835.

Devuyst O, Pattaro C. 2018. The UMOD Locus: Insights into the Pathogenesis and Prognosis of Kidney Disease. J Am Soc Nephrol. 29(3):713–726. doi:10.1681/ASN.2017070716.

Dilimulati K, Orita M, Yonahara Y, Imai FL, Yonezawa N. 2022. Identification of Sperm-Binding Sites in the N-Terminal Domain of Bovine Egg Coat Glycoprotein ZP4. International Journal of Molecular Sciences. 23(2). doi:10.3390/ijms23020762.

Feng J, Tian H, Hu Q-M, Meng Y, Xiao H-B. 2018. Evolution and multiple origins of zona pellucida genes in vertebrates. Biology Open. 7:bio036137. doi:10.1242/bio.036137.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28(23):3150–3152. doi:10.1093/bioinformatics/bts565.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. PNAS. 99(16):10533–10538.

Goudet G, Mugnier S, Callebaut I, Monget P. 2008. Phylogenetic Analysis and Identification of Pseudogenes Reveal a Progressive Loss of Zona Pellucida Genes During Evolution of Vertebrates1. Biology of Reproduction. 78(5):796–806. doi:10.1095/biolreprod.107.064568.

Guo L, Wang S, Li M, Cao Z. 2019. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. BMC Bioinformatics. 20(25):700. doi:10.1186/s12859-019-3275-6.

Hart M, Stover D, Guerra V, V Mozaffari S, Ober C, Mugal C, Kaj I. 2018. Positive selection on human gamete-recognition genes.

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. New York: Springer (Springer Series in Statistics).

Hawkins DM. 2004. The Problem of Overfitting. J Chem Inf Comput Sci. 44(1):1–12. doi:10.1021/ci0342472.

Howe KL, Achuthan P, Allen James, Allen Jamie, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. 2021. Ensembl 2021. Nucleic Acids Research. 49(D1):D884–D891. doi:10.1093/nar/gkaa942.

Jovine L, Qi H, Williams Z, Litscher E, Wassarman PM. 2002. The ZP domain is a conserved module for polymerization of extracellular proteins. Nature Cell Biology. 4(6):457–461. doi:10.1038/ncb802.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021 Jul 15. Highly accurate protein structure prediction with AlphaFold. Nature. doi:10.1038/s41586-021-03819-2. https://doi.org/10.1038/s41586-021-03819-2.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 30(4):772–780. doi:10.1093/molbev/mst010.

Killingbeck EE, Swanson WJ. 2018. Chapter Fourteen - Egg Coat Proteins Across Metazoan Evolution. In: Litscher ES, Wassarman PM, editors. Current Topics in Developmental Biology. Vol. 130. Academic Press. p. 443–488. https://www.sciencedirect.com/science/article/pii/S0070215318300486.

Kim D-K, Kim JA, Park J, Niazi A, Almishaal A, Park S. 2019. The release of surface-anchored α-tectorin, an apical extracellular matrix protein, mediates tectorial membrane organization. Sci Adv. 5(11):eaay6300–eaay6300. doi:10.1126/sciadv.aay6300.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35(21):4453–4455. doi:10.1093/bioinformatics/btz305.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution. 35(6):1547–1549. doi:10.1093/molbev/msy096.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 33(7):1870–1874. doi:10.1093/molbev/msw054.

Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature. 556(7702):452–456. doi:10.1038/s41586-018-0043-0.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658–1659. doi:10.1093/bioinformatics/btl158.

Lin S, Hu Y, Zhu J, Woodruff T, Jardetzky T. 2011. Structure of betaglycan zona pellucida (ZP)-C domain provides insights into ZP-mediated protein polymerization and TGF- binding. Proceedings of The National Academy of Sciences - PNAS. 108:5232–5236. doi:10.1073/pnas.1010689108.

Litscher ES, Wassarman PM. 2020. Zona Pellucida Proteins, Fibrils, and Matrix. Annu Rev Biochem. 89(1):695–715. doi:10.1146/annurev-biochem-011520-105310.

Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P. 2012. Evolution of Genes Involved in Gamete Interaction: Evidence for Positive Selection, Duplications and Losses in Vertebrates. PLOS ONE. 7(9):e44548. doi:10.1371/journal.pone.0044548.

Monne M, Han L, Schwend T, Burendahl S, Jovine L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature. 456(7222):653–7.

Nishimura K, Dioguardi E, Nishio S, Villa A, Han L, Matsuda T, Jovine L. 2019. Molecular basis of egg coat cross-linking sheds light on ZP1-associated female infertility. Nat Commun. 10(1):3086–3086. doi:10.1038/s41467-019-10931-5.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12:2825–2830.

Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36(7):2295–2300. doi:10.1093/nar/gkn072.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 25(13):1605–1612.

Raj I, Sadat Al Hosseini H, Dioguardi E, Nishimura K, Han L, Villa A, de Sanctis D, Jovine L. 2017. Structural Basis of Egg Coat-Sperm Recognition at Fertilization. Cell. 169(7):1315-1326.e17. doi:10.1016/j.cell.2017.05.033.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research. 18(20):6097–6100. doi:10.1093/nar/18.20.6097.

Schrödinger L. 2015. The PyMOL Molecular Graphics System, Version 1.8.

Shu L, Suter MJ-F, Räsänen K. 2015. Evolution of egg coats: linking molecular biology and ecology. Molecular Ecology. 24(16):4052–4073. doi:10.1111/mec.13283.

Sircar A, Chaudhury S, Kilambi KP, Berrondo M, Gray JJ. 2010. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13–19. Proteins: Structure, Function, and Bioinformatics. 78(15):3115–3123. doi:10.1002/prot.22765.

Streuli CH. 2016. Integrins as architects of cell behavior. MBoC. 27(19):2885–2888. doi:10.1091/mbc.E15-06-0369.

Wassarman PM, Litscher ES. 2016. Chapter Thirty-One - A Bespoke Coat for Eggs: Getting Ready for Fertilization. In: Wassarman PM, editor. Current Topics in Developmental Biology. Vol. 117. Academic Press. p. 539–552. https://www.sciencedirect.com/science/article/pii/S0070215315001167.

Weadick CJ. 2020. Molecular Evolutionary Analysis of Nematode Zona Pellucida (ZP) Modules Reveals Disulfide-Bond Reshuffling and Standalone ZP-C Domains. Genome Biology and Evolution. 12(8):1240–1255. doi:10.1093/gbe/evaa095.

Wilburn DB, Swanson WJ. 2017. The “ZP domain” is not one, but likely two independent domains. Molecular Reproduction and Development. 84(4):284–285. doi:10.1002/mrd.22781.

Wilburn DB, Swanson WJ. 2018. Gamete Structure: Egg, Comparative Vertebrate. In: Skinner MK, editor. Encyclopedia of Reproduction (Second Edition). Oxford: Academic Press. p. 204–209. https://www.sciencedirect.com/science/article/pii/B9780128096338205578.

Wong JL, Wessel GM. 2005. Defending the Zygote: Search for the Ancestral Animal Block to Polyspermy. In: Current Topics in Developmental Biology. Vol. 72. Academic Press. p. 1–151. https://www.sciencedirect.com/science/article/pii/S0070215305720019.

Yang KK, Wu Z, Bedbrook CN, Arnold FH. 2018. Learned protein embeddings for machine learning. Bioinformatics. 34(15):2642–2648. doi:10.1093/bioinformatics/bty178.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 24(8):1586–1591. doi:10.1093/molbev/msm088.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Molecular Biology and Evolution. 22(4):1107–1118. doi:10.1093/molbev/msi097.

Zou H, Hastie T. 2005. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 67(2):301–320.