July 3rd -July 5th, 2023

# Protein Evolution Meeting Münster

by

bornberglab.org

## PRESENTATIONS ON MONDAY JULY 3rd, 2023

### Towards artificial proteins and alphabets: lessons from origins-of-life

Klara Hlouchova
*Department of Cell Biology, Faculty of Science, Charles University, Prague, Czech Republic*

All extant cells known to humankind build proteins from the same 20 coded amino acids. The sequence space of the canonical amino acids provides far more options for de novo gene birth and design and is being extensively exploited. Nevertheless, synthetic biology and the study of origins of life imply that functional proteins can be built using smaller and energetically less costly alphabets and that inclusion of non-canonical amino acids can escape the Central Dogma of all life and be of great potential in biotechnology and therapeutic strategies.
While AI-based predictors have learnt to work with the canonical alphabet, the rules guiding the properties of restricted and extended alphabets remain challenging. To approach this, we study highly combinatorial libraries composed of different amino acid repertoires. I will summarize the knowledge we have gained from the study of early origins of life, describe the biophysical properties of sequences with selected non-canonical amino acids and discuss our approaches towards designing novel xeno-alphabets for building artificial proteins.

References:

[1] Makarov M, Sanchez Rocha AC, Krystufek R, Cherepashuk I, Dzmitruk V, Charnavets T, Faustino AM, Lebl M, Fujishima K, Fried SD, Hlouchova K. (2023): *J. Am. Chem. Soc.* 145, 9, 5320–5329.
[2] Heames B, Buchel F, Aubel M, Tretyachenko V, Lange A, Bornberg- Bauer E, Hlouchova K. (2023) *Nature Ecology & Evolution* 10.1038/s41559-023-02010-2
[3] Tretyachenko V, Vymetal J, Neuwirthova T, Vondrasek J, Fujishima K, Hlouchova K. (2022) *Open Biology 12*: 220040.
[4] Giacobelli VG, Fujishima K, Lepsik M, Tretyachenko V, Kadava T, Bednarova L, Novak P, Hlouchova K. (2022) *Mol. Biol.Evol.*, msac032.

### Protein Refoldability Provides Clues Into the Earliest Proteins

Stephen D. Fried*, Philip To, Klara Hlouchova
**Department of Chemistry, Johns Hopkins University, Baltimore, MD, USA*

By coupling structural proteomics with global refolding assays, we have characterized which types of proteins have the capacity to spontaneously refold from a denatured form, which have a reliance on chaperone assistance, and which depend on the ribosome (that is, are obligate co-translational folders).  We can infer that during early protein evolution – prior to the emergence of chaperones or ribosomal translation – refoldability was an essential characteristic.  Hence, these assays provide a “biophysical” signature for the earliest proteins, that can complement sequence analysis and reconstruction.

Several insights emerge from this analysis that will be discussed. The most refoldable classes are small all-β folds (such as OB-folds, SH3, and Ub-like), supporting their placement as early folds. The topologically more sophisticated folds associated with metabolism (TIM barrels and Rossmanns) are less refoldable and more dependent on chaperones. The specialized folds associated with aminoacyl-tRNA synthetases generally are poor refolders and their biogenesis is more dependent on translation itself. These findings support an alternative model in which translation factors were *not* the earliest proteins, but rather co-evolved with the ribosomal translation explaining their "addiction" to it to fold properly. RNA-binding proteins are generally among the most refoldable proteins, especially those associated with the large ribosomal subunit, but not those associated with the small ribosomal subunit. Hence, the large subunit (and peptide catalysis) likely preceded the advent of polynucleotide-to-peptide decoding.

These findings support the existence of an early protein/RNA world in which ancient proteins closely collaborated with RNA without being encoded by it, and call for more work to elucidate mechanisms of early protein propagation and chemical evolution without direct encoding by a polynucleotide template.

---

## Fold evolution before LUCA

Claudia Alvarez Carreno
*The Parker H. Petit Institute for Bioengineering and Bioscience*
*Georgia Institute of Technology, Atlanta, USA*

We analyzed the deep evolutionary history of the translation machinery, which is an ancient molecular fossil and the birthplace of all folded proteins. We demonstrate common origins of some of the simplest, oldest and most common protein folds: the Oligonucleotide/Oligosaccharide-Binding, (OB); the Scr kinase family Homology 3 (SH3); and the cradle loop barrel (CLB). Proteins with OB, SH3 and CLB folds are found in core cellular processes including translation and replication.

The combined data on beta-barrel folds and circularly permuted proteins support a model of creative destruction in which ancestral folds readily beget new folds. Each step of creative destruction is known to be independently accessible and relatively frequent.

Creative destruction explains the facile emergence of complex from simple architectures in a funneled exploration. In creative destruction, new folds emerge from old folds via gene duplication, expression, exploration of new folding landscapes and adaptation.

---

### Patterns of Duplicate Gene Retention Over Different Timescales and With Different Selective Pressures

David A. Liberles
*Center for Computational Genetics and Genomics and Department of Biology, Temple University, Philadelphia, USA*

Gene duplication is an important process for the evolution of novel gene functions. Previously, we presented the first models that enabled the probabilistic inference of evolutionary mechanisms leading to duplicate gene retention, that differentiated nonfunctionalization, neofunctionalization, subfunctionalization, and dosage balance. Independently, in analyzing duplicate gene retention following two rounds of whole genome duplication in Atlantic salmon, we observed that genes retained after the second round of whole genome duplication were approximately equally likely to be retained whether they were from retained from the first round or were not. This was surprising in light of the commonly accepted gene duplicability hypothesis, that some genes are inherently more likely to be retained due to either their function or to the number of interacting partners they have (with a dosage balance mechanism). Here, we present the formal expectations of the gene duplicability hypothesis in the framework generated for analysis of the Atlantic salmon genome. We generated new data from monocot species with multiple rounds of whole genome duplication that can also generate insight into these evolutionary processes. We also present a new modeling framework to analyze gene duplicates on population genetic timescales. Lastly, we show that subfunctionalization for genes under dosage balance constraints involving interacting partners behaves as a non-neutral process. All of this work together gives an enhanced picture of the dynamics of duplicated genes in genomes.

---

### Modeling protein evolution for phylogenetics on the billion-year timescale

Andrew J. Roger*, Edward Susko, Hector Baños, Charley McCarthy, Kelsey Williamson, Thomas Wong and Bui Quang Minh
**Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada*

The most widely used models of protein sequence evolution used for phylogenetic reconstruction employ an empirically estimated general rate matrix of amino acid ‘exchangeabilities’ (e.g., PAM, WAG or LG), a discrete distribution that accounts for different rates at different sites (e.g., Gamma) and assume a single vector of amino acid stationary frequencies for the protein sequences. However, it is well known that use of such single-matrix models for deep phylogenetic problems – trees that span hundreds of millions or billions of years – can lead to artefacts such as long-branch attraction. The development and use of site-heterogeneous profile mixture models together with empirical exchangeability matrices has addressed some of these problems but can still fail to capture important aspects of protein evolution on the billion-year time scale. Here I will introduce several new site-heterogeneous profile mixture models designed to improve their biological realism and

fit-to-data by allowing for group-specific exchangeability matrix estimation, changing amino acid frequencies over the tree and functional shifts in molecules across branches. I will demonstrate their utility by simulations and real data analysis examples. I will also discuss aspects of protein evolution that are still not well-modeled by these approaches and the problems they pose.

---

### Principles of Protein Evolution from a Biophysical Perspective

Richard Goldstein
*Division of Biosciences, University College London, UK*

There is a long history of using thermodynamic principles in evolutionary theory, from the recognition that biomolecules evolve subject to thermodynamic constraints, to the use of diffusion theory to understand changes in gene frequency, to recognising that many fundamental concepts of statistical mechanics have their analogues in evolution. In particular entropy, the number of ways a given state can be realised, is directly analogous to number of sequences corresponding to a given phenotype. This sequence entropy approach predicts the presence of an 'Evolutionary Stokes shift' where proteins adjust to amino acid substitutions by stabilising the new amino acid, with a significant impact on substitution rates. We would expect this phenomenon to be more general, with entropy resulting in stabilisation of the consequences of all forms of evolutionary change. In particular, we can use these ideas to understand how the organism adjusts to the presence of a de novo protein, and how the continued existence of that protein can be facilitated by this effect.

---

### Studies of the influence of thermodynamic stability on the sequence landscapes of proteins from atomistic simulations

Ingemar André,
*Department of Biochemistry and Structural Biology, Lund University, Sweden*

The requirement that proteins must fold into stable three-dimensional structures is one of the major fitness constraints in evolution. The stability of a protein is dependent on highly specific intermolecular interactions at the atomic scale. In our work, we have studied protein evolution using atomistic energy models coupled with population genetics frameworks to investigate how protein energetics shape sequence landscapes. Our results demonstrate that amino acid substitution rates can be largely predicted based on the stability fitness constraint and the structure of the genetic code. We also developed an all-atom simulator of protein evolution and use it to study how protein stability affects the distribution of proposed and accepted mutations, site-specific rates, and the prevalence of correlated amino acid substitutions. Our results demonstrate how the stability fitness constraint is sufficient for sequence covariation to emerge in protein alignments. Models that calculate site-specific rates based on protein structure and a thermodynamic stability model have shown a significant but modest ability to predict empirical site-specific rates calculated from sequence.

We demonstrate that a fundamental reason for this is that empirical site-specific rates are estimated as averages across many different microenvironments in a phylogeny. Finally, we demonstrate how the model used to predict site-specific rates based on stability fitness constraints can be extended to estimate empirical rates from multiple sequence alignments. This model is shown to perform better than standard site-rate inference approaches on realistically simulated protein alignments and at a fraction of the computational cost.

---

### The erroneous and disordered protein sequence space

Agnes Toth-Petrozcy
*Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany*

My lab seeks to understand protein evolution in the light of erroneous protein production with focus on intrinsically disordered and condensate-forming proteins.
We are fascinated by the robustness and plasticity of proteins towards mutations: a single mutation can result in a novel protein function or a damaged protein leading to disease. I will present how evolutionary and structural information can aid clinical variant interpretation (demag.org <http://demag.org/>).
Our main focus is not only on genetic mutations but also phenotypic mutations that are the result of transcriptional and translational errors. They increase protein heterogeneity within an isogenic population that can be the source of innovation and lead to adaptation. We are exploring the role of phenotypic mutations in evolution by combining theoretical modeling, machine learning and experiments.
Most of our understanding about the molecular history of the cell is biased towards well-conserved ordered protein domains. Intrinsically disordered protein regions (IDRs) are largely unexplored due to their low sequence complexity and low conservation. Yet they have pivotal roles in the cell including facilitating the formation of biomolecular condensates. I will present how we have tackled some of the methodological challenges regarding IDRs and condensates (cd-code.org <http://cd-code.org/>) to aid our systematic understanding of protein sequence to function relationships.

---

### The evolutionary road to complexity of earth’s most important $CO_2$ fixing enzyme.

Georg Hochberg
*Max Planck Institute for Terrestrial Microbiology, Marburg, Germany*

Rubisco is the central $CO_2$ fixing enzyme of the Cavin cycle and responsible for the vast majority of all $CO_2$ fixation on our planet today. In plants, Rubisco undergoes an elaborate set of steps involving the sequential action of at least 6 different dedicated folding and assembly chaperones to assemble into its enzymatically active form. This complexity evolved from much simpler Rubisco ancestors that functioned without any of these additional factors. In this talk I will summarize my lab’s work on retracing the evolution of Rubisco’s complex present-day assembly requirements. Using ancestral sequence reconstruction and the resurrection of billion-year-old Rubiscos, we are learning how this crucial enzyme gradually elaborated its structure and assembly mechanism. Some of these elaborations had history-changing effects on Rubisco's catalytic properties, whereas others appear to be evolutionary accidents that simply became impossible to lose. This work is beginning to illuminate key events in Rubisco's history leading up to and following the evolution of oxygenic photosynthesis, one of the most consequential events in the history of life on earth. It also raises the possibility of learning from evolution to re-simplify and improve the assemblies of agriculturally important Rubiscos.

---

### Ancestral proteins as scaffolds for enzyme engineering and evolution

Jose M. Sanchez-Ruiz
*Department of Physical Chemistry (Faculty of Science), University of Granada, Spain*

Plausible approximations to sequences of ancestral proteins (i.e., proteins from extict organisms) can be derived from the known sequences of their modern descendants using suitable analyses.
The possibility of ancestral sequence reconstruction was proposed by Linus Pauling and Emile Zuckerkandl in 1963. While this was only a theoretical possibility in the mid-twentieth century, ancestral sequence reconstruction has recently become a standard approach in molecular evolution studies, as a result of advances in phylogenetics and bioinformatics, together with the availability of very large sequence databases. In the last ⍰25 years, proteins encoded by reconstructed ancestral sequences (i.e., "resurrected" ancestral proteins, in the common jargon of the field) have proven to be very useful tools to address important problems in molecular evolution. Furthermore, we and others have shown that resurrected ancestral proteins may display properties that are desirable in scaffolds for enzyme engineering. These properties include high stability, as well as substrate and catalytic promiscuity, and will be described in some detail inrelation with the generation of de novo enzymes, a major unsolved problem in protein engineering. Finally, I will discuss the unusual properties and biotechnological potential of ancestral TIM-barrels and the use of resurrected ancestral proteins to study viral evolution.

---

### The androgynous enzymes: how dual functionality evolves

Paola Laurino
*Okinawa Institute of Science and Technology Grad University, Japan*

In the last few decades, the utility of sequence data to understand protein evolution and engineering or design new protein function become critical. One of the most used methods for these scopes is Ancestral sequence reconstruction (ASR). ASR is based on multiple sequence alignment of homologous sequences and allows to reconstruct plausible sequences of the extinct ancestors of a modern protein family. ASR turned to be a great tool to understand how protein function emerge and evolve in nature.
Herein I will show how ASR can be used to discover two of strategies that Nature employs to create enzymes with dual functionalities. In the first case we discovered how an extension of

peripheral loops allows the Glycoside Hydrolase family 19 chitinase enzyme to acquire new antifungal activity. Interestingly, this strategy does not imply the reshaping of the catalytic loop but rather the addition of loops remote from the catalytic pocket. In the second case, the acquisition of RNA methyltrasfersase activity for a DNA methyltrasferase is allowed during evolution because of a mild and persistent selection pressure on the substrate binding pocket.

---

## PRESENTATIONS ON TUESDAY JULY 4th, 2023

---

### Protein function from a network perspective

Christine Brun
*TAGC – Inserm- Aix-Marseille Université, Marseille, France*

I will explain how Network Biology brings new insights in the functions of moonlighting proteins, small peptides and host-microbes relationships.

---

### The role of insertions and deletions on the evolution of new protein folds

Evandro Ferrada
*CeMM - Center for Molecular Medicine, Vienna, Austria.*

Empirical evidence suggests that insertions and deletions are a frequent mutational force in the evolution of protein coding genes. However, due to our incomplete knowledge of the sequence-structure map, it is unknown how often indels lead to foldable proteins, or to proteins with new conformations. To explore this question, I present a simple exact model which handles sequences and conformations of variable length. I use this model to define folds as conformations with a common core structure. Then, I explore the frequency with which indels of different lengths and amino acid composition lead to new, foldable conformations. Finally, I test predictions of the model using protein sequence and structure data.

---

### Foldability potential encoded in intergenic regions and de novo gene birth

Anne Lopes
*Institute of Integrative Biology of the Cell, CEA-CNRS-Université Paris-Sud, France*

It is now evident that noncoding regions are extensively expressed and provide a significant source of novel protein products. While most of these products are expected to be short-lived during evolution, some of them, if beneficial, may be fixed and established as novel genes. However, our current understanding of the sequence and structural properties of these unevolved products and the mechanisms governing their production until fixation remains limited. Particularly, the impact of nucleotide composition on the structural properties and evolutionary trajectories of noncoding sequences is still poorly understood. In this study, we investigated the foldability and other sequence properties of peptides encoded in intergenic ORFs (iORFs) across thousands of eukaryotic genomes with GC contents ranging from 18% to 79%. Our findings reveal that depending on the GC content, iORF-encoded peptides, if translated, would exhibit distinct structural properties. This raises questions about the potential transition, if any, and the nature of the transitory stages between noncoding ORFs and established genes in these different organisms. Using phylostratigraphy, de novo gene identification, and ancestral sequence reconstruction, we traced the evolution of several hundred de novo proteins in 16 organisms with varying GC contents. Our results highlight a significant role of GC content and protein foldability in de novo gene birth and more generally, in the establishment of the genetic code.

---

## Protein evolution on fitness landscapes:Tradeoffs, epistasis and irreversibility

Joachim Krug
*Institute for Biological Physics, University of Cologne, Köln, Germany*
*Email: jkrug@uni-koeln.de*

The fitness landscape has served as a powerful metaphor in evolutionary biology for almost a century. At present, spectacular progress in high-throughput techniques provides increasingly detailed glimpses into the structure of the empirical fitness landscapes of single proteins or entire organisms, and this has triggered a resurgence of interest in mathematical landscape models. In the first part of the talk, I briefly describe two recent examples of models that incorporate structural constraints, such as phenotypic tradeoffs [1] or universal epistasis [2], thereby predicting topographies that differ significantly from standard models such as Kauffman's NK-model. In the second part, I use the fitness landscape perspective to explain the dependence of the fitness effects of substitutions on the changing genetic background along an evolutionary trajectory, conceptualized as contingency and entrenchment [3]. In short, these terms refer to the observation that substitutions occurring at a certain point of the trajectory would have been less beneficial if they had occurred earlier, and are more deleterious to revert at later times. I will show that, despite the apparent irreversibility of this phenomenon, contingency and entrenchment arise generically in dynamics that are manifestly reversible, and they are robust with respect to the structure of the underlying fitness landscape. The second part is based on joint work with Lucy Lansch-Justen.

[1] S.G. Das, S.O. Direito, B. Waclaw, R.J. Allen, J. Krug: Predictable properties of fitness landscapes induced by adaptational tradeoffs. eLife 9:e55155 (2020)
[2] K. Crona, J. Krug, M. Srivastava: Geometry of fitness landscapes: peaks, shapes and universal positive epistasis. J. Math. Biol. 86:62 (2023)
[3] P. Shah, D. M. McCandlish, J. B. Plotkin: Contingency and entrenchment in protein evolution under purifying selection. PNAS 112:E3226-E3235 (2015)

---

## The structure of genotype-phenotype maps makes fitness landscapes navigable

Sebastian Ahnert
*Department of Chemical Engineering and Biotechnology ,University of Cambridge, UK*

In the biological literature fitness landscapes are often described in intuitive terms, as lowdimensional landscapes of the kind encountered in everyday experience with ‘peaks’ and ‘valleys’. Genotype space, however, is extremely high dimensional, resulting in counter-intuitive structural properties of genotype-phenotype maps. We present results showing that these structural properties, such as the existence of large neutral networks, make fitness landscapes navigable. For three biologically realistic models of genotype-phenotype maps—RNA secondary structure, protein tertiary structure and protein complexe—we find that fitness maxima can be reached from almost any phenotype without passing through fitness valleys, even if phenotypes are assigned random fitnesses. This in turn indicates that fitness landscapes only contain very few true valleys. We also consider evolutionary simulations between pairs of functional RNA sequences and show that paths of monotonically increasing fitness are also likely to be used under evolutionary dynamics. Our findings can inform evolutionary predictions as well as directed evolution

---

## Evolutionary landscapes of proteins: known territories and dark matter in sequence space

Erich  Bornberg-Bauer
*Institute for Evolution and Biodiversity, University of Muenster, Germany*

Proteins are the univresal workhorses of all living beings and their evolution reflects the evolution of life since the presumed RNA world. Many mechanisms such as duplication followed by sub- or neo-functionalisation or dosage compensation, epistatic interaction or modular rearrangements have been proposed over the last decades and found strong support from theory and experiments. However, all these mechanisms seem to consider only a small fraction of sequences space occupied by dense clusters amounting to protein families while the overwhelming vastness of sequence space is uncharted  territory.
We use model protein simulations, data analysis and experiments to understand how much of sequence space has been explored by evolution, why some regions are not represented and what the biophysical differences between these fractions are.

---

### Protein evolution from non-genic sequences and existing proteins

Bharat Ravi
*Institute for Evolution and Biodiversity, University of Muenster, Germany*

In the first part of the presentation, I will talk about how new protein coding genes evolve from non-genic sequences. This phenomenon is called de novo emergence. I explore this process using mathematical models and address some open questions in the field. One such question is whether transcription emerges before the emergence of an open reading frame. Using the model, I also ask whether an open reading frame is more easily lost than a transcript of similar length. Higher than expected loss rates could indicate negative selection. In the second part of my presentation, I talk about how the chaperones GroEL, Hsp90, and DnaK influence directed evolution of green fluorescent protein. Chaperones help protein fold and can rescue the deleterious effects of destabilizing mutations. This in turn can provide phenotypic robustness against mutations, and can increase the genetic diversity of protein variants under stabilizing or directional selection. To this end, I analyse how the phenotype and genotypes of evolving GFP variants is influenced when these chaperones are transiently overexpressed during selection. My study shows that the three different chaperones affect protein evolution differently and chaperone overexpression can sometimes be counterproductive to mutational robustness.

---

### Understanding the origins and long-term evolutionary trends of novel genes in the entire budding yeast subphylum

Nikolaos Vakirlis
*Institute For Fundamental Biomedical Research, B.S.R.C. “Alexander Fleming”, Vari, Greece.*

Novel genes can emerge entirely de novo, are important drivers of adaptation and often underlie species-specific characteristics. Here, we conducted the largest scale computational investigation of de novo emerged novel genes to date, exploiting a rich dataset comprised from 332 budding yeast genomes, spanning the entire biodiversity of the Saccharomycotina subphylum. We were able to identify over 400,000 taxonomically restricted genes (TRGs) at different taxonomic levels, from species-specific ones to conserved across yeasts. This enabled us to reveal macro-evolutionary trends of gene and protein properties that hold across yeast lineages, including that the GC content of genes does not change with age while intrinsic protein disorder consistently decreases. By employing synteny analysis, we isolated more than 10,000 de novo genes, predicted their three-dimensional structure and found cases of potential convergence towards existing folds. Additionally, we found thousands of TRGs that have diverged beyond recognition and have properties contrasting those of de novo genes such as longer length and lower biosynthetic cost. We then investigated the cryptic property of intergenic regions to encode transmembrane domains, if theoretically translated, more frequently than expected by chance, a finding previously reported in baker’s yeast. We found that this transmembrane domain-forming enrichment is present in the genomes of almost all yeast species and is not explained by either the hydrophobic content or their size of the sequences. Finally, we uncovered a correlation, across species, between this intergenic enrichment and the number of transmembrane domains in evolutionarily young genes, hinting towards a link to de novo emergence.

---

### The Vast and Evolutionarily Novel Yeast Translatome

Anne-Ruxandra Carvunis
*Computational and Systems Biology, School of Medicine, Universtiy of Pittsburgh, USA*

The yeast genome contains more than 200,000 open reading frames (ORFs), of which ~5,000 are annotated with confidence as protein-coding genes. Most yeast ORFs are short and lack evidence of conservation between closely related species or between isolates. These sequence properties are consistent with the notion that these ORFs transiently emerge and then disappear with random mutations in the absence of selection. As a result, these ORFs remain unannotated. However, we show by integrative ribosome profiling analyses that up to 20,000 unannotated yeast ORFs are translated. Through systematic analyses of gain- and loss-of-function mutants, high-content microscopy and integrative transcriptomics, we are uncovering evidence that transiently emerging ORFs carry a surprisingly profound impact on cell biology and fitness despite their seemingly random evolutionary patterns. We discuss the implications of our findings for the systems biology of evolutionary innovation.

---

### Post-endosymbiotic innovations for protein import and regulation of gene expression in mitochondria and chloroplast

Ingrid Lafontaine, *Professeure Sorbonne Université*
*Biologie du Chloroplaste et Perception de la Lumière chez les Microalgues (UMR7141, CNRS-SU),Institut de Biologie Physico-Chimique, Paris, France*

Endosymbiosis that gave rise to extant mitochondria and chloroplasts produced a massive gene transfer from the organelle progenitors to the nucleus of the host cell. Most of the organelle proteome is now nuclear-encoded, translated in the cytosol and imported back in the organelle, that raises two major evolutionary questions:
i) how these imported proteins acquired an organelle targeting sequence and how the organelle progenitors acquired the corresponding import machinery?
ii) how is coordinated the expression of organelle genome with the nuclear genome, required for cell viability and acclimation of the organelle activity?
To answer to the first question, we propose an antimicrobial origin of organelle targeting peptides. In support to this hypothesis, I will present recent results providing conclusive computational and functional evidence for an evolutionary link between targeting peptides and a subtype of ribosomally produced antimicrobial peptides that share a propensity to form amphipathic alpha-helices (HA-RAMPs) when in contact with a membrane.

Regarding the second question, the chloroplast and the mitochondrial genomes are mainly regulated post-transcriptionally, by RNA-binding proteins encoded in the nuclear genome. The evolution of these RNA-binding proteins, together with their network of RNA targets, conditions the response of an eukaryotic organism to environmental changes. This knowledge is based on pioneering studies in photosynthetic eukaryotes of the green lineage (derived from an endosymbiotic event involving a heterotrophic protist and a cyanobacteria), but nothing is known about the regulation of the chloroplast in most of the photosynthetic eukaryotes, comprising a large proportion of organisms, such as Stramenopiles, derived from a secondary endosymbiosis, involving the capture of at least a red alga by another eukaryote. I will present our approach to identify protein candidates for the regulation of chloroplast gene expression at the level of all the photosynthetic eukaryotes.

---

**Evolution of new peptides from non-coding small Open reading Frames**

Juan Pablo Couso*, Pedro Patraquim, Casimiro Baena, Eugenio Mangas, Ana Platero, Emile Magny

**Centro Andaluz de Biología del Desarrollo, Universidad Pablo de Olavide, Sevilla, Spain*

Small, or short, ORFs are DNA and RNA sequences that could be translated into peptides or short proteins of less than 100 amino-acids. Hundreds of thousands of smORF sequences are found in eukaryotic genomes, and thousands can be mapped to transcripts, in many cases, to putative non-coding RNAs. smORFs have remained uncharacterised and even non-annotated in most genomes due to a mixture of technical difficulties (due to their short size), and intellectual prejudice. Over a number of years, we have assessed smORF conservation, (using custom bioinformatic methods), translation, (using extensive ribosomal profiling and proteomics), and function (using CRISPR genetics). We propose that smORFs belong to different classes according to their RNA features and AA sequence, which correlate with their general function and conservation, from inert DNA sequences to transcribed ones, and finally, to sequences producing peptides able to directly regulate canonical proteins. Further, we propose that smORFs arise at random in non-coding regions of genomes and RNAs, and that smORF classes actually represent stages in their evolution towards novel protein coding sequences. This is supported by our discovery of a positive correlation between depth of conservation and translation levels, and by the prevalence of natural selection amongst evolutionary recent, yet translated, smORFs. Finally, we observe that coding features seem to be independently acquired in a ´mosaic´ fashion.

---

## Origin, Evolution, and Function of Orphan genes in *Pristionchus* Nematodes

Christian Rödelsberger
*Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Tübingen, Germany*

Nematodes are one of the most diverse and species-rich animal phyla. Their small size and short generation time makes free-living nematodes excellent model organisms with *Caenorhabditis elegans* being the prime example. More than two decades ago, the nematode *Pristionchus pacificus* was introduced as a model for comparative studies with *C. elegans*. When its genome was sequenced, around one third of genes were classified as orphan genes based on the lack of homologs in other nematodes. Over the past years, we discovered through unbiased genetic screens that orphan genes are central for developmental decisions and for the evolution of novel behaviors. Furthermore, we characterized the origin and evolution of *P. pacificus* orphan genes using deep taxon phylogenomics. Here, I will summarize our recent results about their evolutionary dynamics and then present our latest efforts in linking orphan genes with biological functions. Specifically, from transcriptomic data of worms grown on different bacteria, we generated coexpression networks that harbor thousands of environmentally responsive orphan genes. Inferring cofunctionality based on coexpression, we were able to associate the vast majority of these orphan genes with biological processes and tissues. This confirmed previously observed associations between orphan genes and spermatogenesis, but also revealed potential roles of orphan genes in the nervous system and metabolism. Furthermore, our work demonstrates that novel genes are not randomly attached to existing modules, as certain modules are much more likely to integrate novel genes than others. This links the emergence of orphan genes with the evolution of biological networks and environmental regulation.

## PRESENTATIONS ON WEDNESDAY JULY 5th, 2023

### The origin and evolution of de novo genes in *Drosophila*

Li Zhao*, Junhui Peng, Evan Witt, Nicolas Svetec
**Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University*

Understanding the origin and evolution of genes is crucial to explaining the origin and evolution of novel phenotypes and ultimately the diversity of life. However, the mechanisms through which novel genes originate and are co-opted into functional networks remain largely unknown. Recent work using genomics, next-generation sequencing, and population genomics has revealed that de novo genes—genes born from ancestrally non-genic sequences—contribute to gene innovation. However, the earliest steps in the birth process of de novo genes and how de novo genes are maintained in populations and species were largely unknown. We combined high-quality base-level whole genome alignments, bioinformatics analysis, and computational structure modeling to study the origination, evolution, and protein structure of lineage-specific de novo genes. We observed a gradual shift in sequence composition, evolutionary rates, and expression patterns based on gene ages. Intriguingly, we noted minimal protein structural changes in Drosophilinae lineage de novo genes. We identified several de novo gene candidates with potentially well-folded protein products, many enriched with transmembrane and signal proteins compared to other protein-coding genes. Ancestral sequence reconstruction revealed that these proteins are often born folded. Our study provides a systematic overview of the origin, evolution, and structural changes of Drosophilinae-specific de novo genes.

### The function of random peptides and mouse de novo genes in a human cellline

Diethard Tautz*, Silvia Aldrovandi, Johana Castro Fajardo,
**Max-Planck Institute for Evolutionary Biology, Plön, Germany*
*#tautz@evolbio.mpg.de*

New genes can evolve through duplicaAon and divergence from exisAng genes, or de novo out of non-coding sequences. While de novo evoluAon of genes was contenAous for some Ame, it is now supported by ample evidence, including funcAonal studies, as well as inferences drawn from expressing random coding sequences in bacteria and plants.

We have chosen here a combinaAon of previous approaches to study de novo evolved genes in a human cell line. First, we show that expression of a library of random sequences from a chromosomal integraAon site leads to differenAal fitness effects in the cells that harbor them. In a second experiment, we ask whether genes that have evolved de novo in the mouse lineage have specific effects in human cells when they are over-expressed. This approach addresses the quesAon of whether a de novo evolved protein can interact with a molecular network in a species in which it has not evolved and could therefore not have been adapted for funcAon. This is somewhat akin to expressing random sequences, but pre-selected for sequence variants that were already evoluAonarily tested in another lineage. By studying the transcriptomic response to the expression of the mouse de novo genes, we find differenal and overlapping effects for each of the mouse genes.

---

### Exploring new sequence space with de novo protein design and generative language models

Sergio Romero-Romero
*Department of Biochemistry, University of Bayreuth, Germany*

The ability to create and engineer stable proteins with custom-made functions is a major goal in biochemistry, which has been addressed through protein design, expanding our knowledge about the interconnection among protein sequence, structure and stability. In this talk, I will discuss two different strategies to explore new protein sequence space and characterize the biophysical properties of these newly generated macromolecules. In the first part, I will present a computational strategy to design collection of stable de novo TIM barrels (DeNovoTIMs) using a computational fixed-backbone and modular approach based on improved hydrophobic packing of this protein fold. DeNovoTIMs navigate a region of the stability landscape previously uncharted by natural proteins. Significant non-additive or epistatic effects were observed when stabilizing mutations from different regions of the barrel were combined and the molecular basis of epistasis in these new proteins appears to be related to the extension of the hydrophobic cores. In the second part, I will show how recent progress in transformer-based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Motivated by this success, we explored new sequence space using two generative language models, ProtGPT2 and ZymCTRL, models trained on natural proteins that generates new protein sequences following the principles of natural ones. Experimental validation of a set of novel proteins show that many of these are well expressed in bacterial host cells and contain the ability to fold and function as their natural counterparts, highlighting the potential of such approaches for protein design. Both strategies show how designing novel macromolecules by de novo protein design and generative language models provide crucial information on the molecular determinants that modulate structure and stability, therefore, an important step towards the fine-tuned modulation of protein function by design.

---

### Proteins with multiple G protein-coupled receptor domains

Nir Ben-Tal
*The George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel*

Currently known G protein-coupled receptors (GPCRs) have a single transmembrane domain. Many GPCRs form dimers that have two transmembrane domains (one per protein), and there are indications that this dimeric interaction is functionally meaningful. Here, based on sequence analysis and structure predictions, we report the existence of 57 proteins with two, three, or four GPCR domains within the same protein chain. We analyze the structures of these multi-GPCRs and show that almost all have DRY/NPxxY motifs, a strong indication of signaling activity. By homology, most of the multi-GPCRs that we identified are olfactory-related; a few are chemokine-related. Multi-GPCR candidates are found in various Chordata species including fish, camel, marmite, Chinese hamster, and new world monkeys. The discovery of receptors with multiple transmembrane domains suggests the possibility for signal regulation and amplification within an individual receptor, revealing another step in GPCR evolution and a new layer of complexity in signal transduction.

---

### Retracing the Rapid Evolution of an Herbicide-Degrading Enzyme by Protein Engineering

Reinhard Sterner*, *Markus R. Busch, Lukas Drexler, Dhani Ram Mahato, Caroline Hiefinger, Sílvia Osuna*

**Institute of Biophysics and Physical Biochemistry, Regensburg Center for Biochemistry, University of Regensburg, D-93040 Regensburg, Germany.*

The mechanisms underlying the rapid evolution of novel enzymatic activities from promiscuous side activities are poorly understood. Recently emerged enzymes catalyzing the catabolic degradation of xenobiotic substances that have been introduced into the environment only during the last decades provide an exquisite opportunity to study these mechanisms. A prominent example is the herbicide atrazine (2-chloro-4-ethylamino-6-isopropylamino-1,3,5-triazine), which is degraded by enzymes constituting the Atz pathway. Here, we analyzed the evolution of the hydroxyatrazine ethylaminohydrolase AtzB, a Zn(II)-dependent metalloenzyme that adopts the popular amidohydrolase fold and catalyzes the second step of the Atz pathway. We searched for promiscuous side activities of AtzB, which might point to the identity of its progenitor. These investigations revealed that AtzB has low promiscuous guanine deaminase activity. Furthermore, we found that the two closest AtzB homologues, which have not been functionally annotated up to now, are guanine deaminases with modest promiscuous hydroxyatrazine hydrolase activity. Based on sequence comparisons with the closest AtzB homologues, the guanine deaminase activity of AtzB could be increased by three orders of magnitude by only four active site mutations. Interestingly, the inverse four mutations introduced into AtzB homologues significantly enhanced and in one case even surpassed the hydroxyatrazine hydrolase activity of wild-type AtzB. Extensive molecular dynamics simulations combined with docking studies elucidated the structural and molecular basis for the mutation-induced activity changes. The

example of AtzB highlights that novel enzymes with high catalytic proficiency can evolve from low promiscuous side activities within a short period of time by only few mutational events.

---

### Sequence space and protein functional classification

Alessandra Carbone
*Department of Computational and Quantitative Biology, Sorbonne Université-CNRS, Paris*

Functional classification of proteins from sequences alone has become a critical bottleneck in understanding the myriad of protein sequences that accumulate in our databases. The great diversity of homologous sequences hides, in many cases, a variety of functional activities that cannot be anticipated. Their identification appears critical for a fundamental understanding of the evolution of living organisms and for biotechnological applications.

In this talk, I will explain how rethinking the sequence space with multiple profile models leads to the functional classification of proteins. I will present ProfileView, a sequence-based computational method designed to functionally classify sets of homologous sequences. ProfileView classifies protein families by enriching known functional groups with new sequences and discovering new groups and subgroups.

---

### On the Structure and Evolution of Prokaryotic Surface-Layer Proteins

Vikram Alva
*Department of Protein Evolution, Max Planck Institute for Biology Tübingen, Germany*

The outermost surface of many archaea and bacteria is covered with a surface-layer (S-layer), a two-dimensional paracrystalline lattice formed through the self-assembly of monomeric proteins. S-layers serve crucial roles in cellular functions, including cell shape maintenance, cellular defense, biomineralization, cell adhesion, and molecular recognition. Until recently, information concerning their evolutionary and structural biology aspects was limited. In my presentation, I will discuss the latest advancements in understanding S-layer proteins, emphasizing the common principles underlying their self-assembly and the insights they offer into the evolution of molecular recognition in eukaryotes.

---

### The Protein Universe Atlas: protein family discovery at unprecedented scales

Joana Pereira
*Computational Structural Biology, Biozentrum, University of Basel*

The “protein universe” refers to the collection of all natural protein sequences. In this conceptual representation, protein families and superfamilies are surrounded by dark areas seemingly unexplored by Nature. Due to large-scale genomic projects there is an exponential increase in protein sequence information deposited and automatically annotated on the basis of homology to proteins of known function. However, the number of “hypothetical proteins” and “proteins of unknown function” is increasing proportionally. This can be due to the low sensitivity of the methods behind these efforts, but also due to the presence of sequences belonging to not hitherto described biological systems that populate “dark areas of the protein universe”.

We have recently revised this hypothesis and found that about 10% of all sequences in UniProt, covering ~35% of all UniRef50 clusters, correspond to seemingly functionally dark sequences. Using large-scale protein similarity approaches, we modelled the landscape of those proteins that can be modelled in 3D by AlphaFold at high predicted. Our results indicate that dark proteins complement the landscape occupied by those with at least one well annotated homolog (i.e., bright proteins) and while most form bridges between well studied areas of protein sequence space, some form unique, well-delimited clusters. These correspond to independent novel families unrelated to any other previously described. Our results highlight that as a community we are closer to unravelling the hidden mysteries of the Protein Universe, from unknown biological systems to new technological applications, and we make our results freely available as an online, interactive webresource, which allows for the exploration of protein sequence space with a few clicks – The Protein Universe Atlas.

---

### Bridging themes: similar protein segments shared between domains of different evolutionary lineages

Rachel Kolodny
*Department of Computer Science, University of Haifa, Israel*

The emergence of novel proteins, beyond these that can be readily made by duplication and recombination of preexisting domains, is elusive. The vast majority of theoretically possible polypeptide chains do not fold, let alone confer function. Hence, just like in protein engineering, protein evolution from preexisting building blocks has clear potential advantages over ab initio emergence from random sequences. Short themes which are segments, or fragments of intact domains, are particularly interesting because they may provide hints about the emergence of domains. Likewise, these themes may be used in engineering novel domains. We computationally search for traces of such ancient reuse events, that can be found within contemporary proteins in the form of reused segments at the sub-domain level. To this end, we curated datasets of short themes, of 20-80 residues, that are unexpectedly shared between domains that are considered to have emerged independently. We will describe two complementary efforts, each with approximately 500 bridging themes, shared between folds, many of them ancient, from across protein space. These include the Rossmanns, P-loop NTPase, TIM-barrel, Flavodoxin, Ferredoxin- like, Rossmann/Trefoil-Plexin like, and Beta-propellers. Interestingly, despite sharing high sequence similarity, the bridging themes often do not retain their structure. This may be because conformation flexibility may confer an evolutionary advantage, in that it fits in multiple environments. We will describe our successful search strategies, and several interesting bridging themes examples.

References:

Qiu, K., Ben-Tal, N., Kolodny, R. Similar protein segments shared between domains of different evolutionary lineages, Protein Science (2022)

Longo, L. , Kolodny, R. , McGlynn SE. , Evidence for the Emergence of β-Trefoils by 'Peptide Budding' from an IgG-like β-Sandwich, PloS Computational Biology (2022)

Kolodny, R. , Nepomnyachiy, S., Tawfik, D. S., Ben-Tal, N. Bridging themes: short protein segments found in different architectures, Molecular Biology & Evolution, 38 (6), 2191-2208 (2021)

---

### OpenFold: Lessons learned and insights gained from rebuilding and retraining AlphaFold2

Mohammed AlQuraishi
*Department Systems Biology, Columbia University, New York, USA*

Talk Abstract: AlphaFold2 revolutionized structural biology by accurately predicting protein structures from sequence. Its implementation however (i) lacks the code and data required to train models for new tasks, such as predicting alternate protein conformations or antibody structures, (ii) is unoptimized for commercially available computing hardware, making large-scale prediction campaigns impractical, and (iii) remains poorly understood with respect to how training data and regimen influence accuracy. Here we report OpenFold, an optimized and trainable version of AlphaFold2. We train OpenFold from scratch and demonstrate that it fully reproduces AlphaFold2's accuracy. By analyzing OpenFold training, we find new relationships between data size/diversity and prediction accuracy and gain insights into how OpenFold learns to fold proteins during its training process.

---

## Exploring Protein Evolutionary Space with Deep Learning

Christine Orengo
*Structural & Molecular Biology Div of Biosciences, University College, London, UK*

Powerful new AI and Deep Learning (DL) based tools are helping to transform our understanding of the protein landscape. Thanks to DeepMind and META we now have hundreds of millions of good quality 3D protein structure models available. The Orengo group have developed DL tools to detect evolutionary relationships between these structures and to explore the number of evolutionary families present in 21 model organisms. A relatively small number of novel families have been identified but very few new 3D architectural arrangements. Our CATH domain structure classification is being used to classify this data and we are also exploiting DL and more classical structural analyses to explore functional relationships between the evolutionary relatives within the CATH superfamilies. An example of analyses between sub-families in the highly functionally diverse HUP superfamily will be presented.

## POSTER SESSION ON MONDAY July 3$^{rd}$ ,23

### Experimental characterisation of de novo proteins and their unevolved random-sequence counterparts

Margaux Aubel*, Brennen Heames, Filip Buchel, Vyacheslav Tretyachenko, Dmitry Loginov, Petr Novák, Andreas Lange, Erich Bornberg-Bauer & Klára Hlouchová
**Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

De novo gene emergence provides a route for new proteins to be formed from previously non-coding DNA. Proteins born this way are considered random sequences, and typically assumed to lack defined structure. While it remains unclear how likely a de novo protein is to assume a soluble and stable tertiary structure, intersecting evidence from random-sequence and de novo- designed proteins suggests that native-like biophysical properties are abundant in sequence space.
Taking putative de novo proteins identified in human and fly, we aim to experimentally characterise libraries of these sequences to assess their solubility and structure propensity. We compare the de novo protein sequences to a set of synthetic random proteins with no evolutionary history. Bioinformatic prediction suggests that de novo proteins may have remarkably similar distributions of biophysical properties to unevolved random sequences of a given length and amino acid composition. However, upon expression in vitro, de novo proteins exhibit higher solubility which is further induced by the DnaK chaperone system. We suggest that while synthetic random sequences are a useful proxy for de novo proteins in terms of structure propensity, de novo proteins may be better integrated in the cellular system than random expectation, given their higher solubility.

### Emergence of the CatSpermasome

Alina Berger
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Research on male infertility has become increasingly important over the past years as overall fertility is declining world wide. Examples of events leading to male infertility are detrimental deformations of sperm or mutations within sperm specific proteins. Sperm specific ion channels (SSICs) are important proteins for the functionality of sperm, and mutations within SSICs could lead to reduced fertility or infertility. The CatSpermasome is a complex calcium SSIC located at the principal piece of the sperm tail, which is essential for successful fertilization. The core of the human CatSpermasome comprises four subunits which

constitute the channel pore. Until today 14 subunits of this ion channel are known in total. On the basis of the recently published CatSpermasome structure from mice, extensive genomic data from human populations and many vertebrate genomes from fish to apes, studying its emergence and its influence on male fertility can be performed.
The overarching aim is to trace the evolution of functional subunit interplay and detect selection and protein sequence signatures which could be related to sperm malfunction. Here, I show the acquisition of all known subunits of the CatSpermasome across the vertebrate and an ape phylogeny. (1)
I study the selection acting on single subunits of the CatSpermasome (2) per residue and (3) per branch in the phylogenetic trees. The selection per residue (2) is shown for the core of the human CatSpermasome predicted by AlphaFold2. For the α I subunit, sites that are selected positively in apes are also found to be selected positively in vertebrates. Preliminary results for the largeextracellular beta subunit indicate positive selection in some branches (3).

### Tracing and reconstructing protein evolution with protein domains

Abdulbaki Coban, Elias Dohmen, Carsten Kemena
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Domains are the functional, independently folding building blocks of proteins. Their high conservation, even across kingdoms, makes them suitable molecular markers for long evolutionary time scales. Functional innovation in protein evolution is mainly generated through rearrangements of domains into novel combinations. We can trace protein evolution and such innovation by reconstructing ancestral domain contents and combinations and follow rearrangements through time or across phylogenetic trees.

DomainWorld (https://domainworld.uni-muenster.de/) is a software suite providing bioinformatic tools to facilitate analyses of protein evolution based on domains. The tools provide functionality in a broad range from early steps such as annotating amino acid sequences with protein domains (RADIANT ) to full domain-based analyses including reconstruction of ancestral domain contents and tracing domain rearrangements over evolutionary time (DomRates). With the example of DomRates we present how we can utilise these tools to gain a deeper understanding of protein evolution and identify evolutionary paths towards novel protein functions as well as lost functionality at certain evolutionary time points based on domains as modular molecular markers.

## Theoretical models in evolution: modeling de novo transcript and transposable element dynamics

Pete Czuppon

*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

I present three applications of theoretical models in evolutionary biology that my research is focused on at the moment: (i) estimation of birth and death rates of de-novo transcripts, (ii) prediction of transposable element abundance in genomes, and (iii) optimization of drug dosage to treat bacterial infections. (i) Some non-genic transcripts arise from previously non-transcribed genomic regions and are therefore referred to as de novo transcripts. The protein-coding potential of these de novo transcripts has sparked the interest of researchers to uncover and understand the early stages of a novel gene. Based on the transcript frequency spectrum derived from different isolines of Drosophila melanogaster, we estimate de novo transcript birth and death rates. (ii) Transposable elements (TEs) are self-replicating genomic elements that populate many species genomes. Here, we study the interplay of an autonomous/non-autonomous TE pair. Non-autonomous TEs rely on the transposition machinery of autonomous TEs and through that naturally limit the spread of autonomous TEs independent of host control. We aim at distinguishing between self-regulatory and host-control mechanisms from data. (iii) The dosage of an antibiotic drug plays an important role for the evolution of antibiotic resistance. We mathematically derive the concentration that increases the risk of resistance evolution the most, which in turn gives insight into optimal treatment concentrations.

---

## Probing the evolvability of ancestral and extant sulfatases using fluorescence-activated droplet sorting (FADS)

Berndjan Eenik

*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Directed Evolution is a useful method in protein engineering as it allows the rapid screening of large protein libraries for variants with improved activity towards desired reactions. With the use of Ancestral Sequence Reconstruction (ASR) it is possible to infer the DNA and protein sequences of ancestral enzymes representing the branching points in a phylogenetic tree. Although these ancestral sequences can no longer be found in nature, studying these enzymes may provide insight into protein evolution and can be provide an additional repertoire of enzymatic function. We use the alkaline phosphatase (AP) superfamily model system to compare the evolvability of ancestral and extant enzymes in the same superfamily. To facilitate the screening of large directed evolution libraries we use Fluorescence Activated Droplet Sorting (FADS). Individual cells expressing a unique library variant were inoculated into picolitre volume droplets and incubated in order to obtain duplicates of each unique library variant for improved recovery. Different methods of sorting both individual cells and cells grown inside a droplet were compared for the high-throughput sorting of intact cells displaying enzyme on the outer membrane using the E. coli autotransporter system. It was found that cell growth inside a droplet prior to sorting increased both the recovery and percentage of true positive variants. Using the improved method sixteen libraries of extant and ancestral libraries with distinct mutational loads were created. We found that increasing the mutational load resulted in a higher proportion of greatly improved variants but a lesser total number of improved variants. Furthermore, we found that effects of mutational load differed quantitatively between enzymes. Although ancestral enzymes did not show the highest proportion of active variants, ancestral enzymes did show the highest proportional increases in activity.

---

## Challenges in Structure Prediction of *de novo* Emerged and Random-sequence Proteins

Lars Eicholt, Margaux Aubel, Lasse Middendorf, Erich Bornberg-Bauer
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

*De novo* protein coding genes emerge from scratch in the non-coding regions of the genome and have, per definition, no homology to other genes. Therefore, their encoded *de novo* proteins belong to the so-called "dark protein space". So far, only four *de novo* protein structures have been experimentally approximated. Low homology, presumed high disorder and limited structures result in low confidence structural predictions for *de novo* proteins in most cases. Additionally, low confidence predictions have widely been associated with high intrinsic disorder.

Here, we look at the most widely used structure and disorder predictors and assess their applicability for *de novo* emerged proteins and correlations between disorder and low confidence predictions. Since AlphaFold2 is based on the generation of multiple sequence alignments and was trained on solved structures of largely conserved and globular proteins, its performance on *de novo* proteins remains unknown. More recently, natural language models of proteins have been used for alignment-free structure predictions, potentially making them more suitable for *de novo* proteins than AlphaFold2.

We suggest that, while in some cases protein language model based approaches might be more accurate than AlphaFold2, the structure prediction of *de novo* emerged proteins remains a difficult task for any predictor, be it disorder or structure. Especially, correlations between low confidence predictions and disorder are different for random and *de novo* proteins, and differ from conserved ones.

---

## Double-modified AdoMet analogues for selective targeting of methyltransferases

M. Erguven*, Aileen Peters, Ann-Marie Lawrence, E. Karaca, A. Rentmeister,
*Department of Chemistry and Pharmacy, Institute of Biochemistry, University of Münster*

Methyltransferase-based bioalkylation has been an important tool in the field of biomolecular labeling. We have previously deployed one-pot biocatalytic cascades by combining methionine adenosyltransferases (MATs) and methyltransferases (MTases) together, as a means to transfer clickable or photocaging groups to various substrates[1–4]. However, MTase promiscuity poses a problem for such applications in living systems. For this reason, we have recently focused on base modifications of the MTase cosubstrate, AdoMet, aiming to achieve selective labeling[5]. In the present study, we further explored this option by expanding our set of ATP analogues. We used a thermostable MAT from *Methanocaldococcus jannaschii* (PC-MjMAT) that is highly active at 37°C and can accept bulky *N*$^6$ base modifications, effectively producing the substrate AdoMet analogues for the MTases. These MAT reactions were combined in cascades with three different MTases, named NovO, RnCOMT, and GlaTgs2. By using bulky *N*$^6$ base-modified ATP analogues as the starting material, we achieved MTase selectivity to a degree *in vitro*. We believe that our results will encourage further research aiming to achieve orthogonality in MTase-based biomolecular labeling.

Literature:

[1] Holstein, J. M. et al. A. Dual 5′ Cap Labeling Based on Regioselective RNA Methyltransferases and Bioorthogonal Reactions. *Chemistry – A European Journal* 2017, *23* (25), 6165–6173.

[2] Michailidou, F. et al. Engineered SAM Synthetases for Enzymatic Generation of AdoMet Analogs with Photocaging Groups and Reversible DNA Modification in Cascade Reactions. *Angew. Chem. Int. Ed.* 2021, *60* (1), 480–485.

[3] Muttach, F. and Rentmeister, A. A Biocatalytic Cascade for Versatile One-Pot Modification of MRNA Starting from Methionine Analogues. *Angew. Chem. Int. Ed.* 2016, *55* (5), 1917–1920.

[4] Peters, A. et al. Visible-Light Removable Photocaging Groups Accepted by MjMAT Variant: Structural Basis and Compatibility with DNA and RNA Methyltransferases. *ChemBioChem* 2022, *23* (1).

[5] Cornelissen, N. V. et al. Rentmeister, A. Nucleoside-Modified AdoMet Analogues for Differential Methyltransferase Targeting. *Chem. Commun.* 2020, *56* (14), 2115–2118.

## Transposable Elements influence new transcripts emergence

Anna Grandchamp, Marie Kristin Lebherz, Bertrand Fouks, Erich Bornber-Bauer,
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Abstract 1.

De novo genes are genes that emerge from noncoding regions of a genome via a succession of mutations. Through these mutations, two main steps have to be achieved: the emergence of an open reading frame (ORF) and the emergence of a transcription event. Although the mechanisms enabling the emergence of an ORF are well documented, little is known about the mechanisms enabling new transcription events.

We searched for de novo transcripts by using newly assembled genomes and transcriptomes of seven lines of Drosophila melanogaster originating from European and African populations. We studied motifs composition and transposable elements upstream intergenic de novo transcripts to understand their possible mechanisms of emergence.

Up- and downstream regions of intergenic de novo transcripts overlapped more significantly with TEs and contained higher cpg content than random intergenic region. TEs overlapping with de novo transcripts belonged to different families, with a predominancy of LTR elements from the gypsy family. Most TEs overlapping with de novo transcripts were not present in homologous untranscribed sequences from outgroup, suggesting that several TEs may still be actively transposing, and that this transposition impacted transcripts emergence. lines Moreover, low identity enhancers motifs were enriched upstream de novo transcripts compared to other regions, with highest enrichment upstream de novo transcripts overlapping with TEs.

Taken together, our results suggest that TE insertions produce a mutational hotspot promoting transcription initiation. Overall, our study highlights the importance of TEs in the emergence of new transcripts.

---

## High turnover of de novo transcripts in Drosophila melanogaster

Anna Grandchamp, Pete Czuppon, Erich Bornber-Bauer,
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

De novo transcripts are non-genic transcripts that are newly transcribed in species or populations. Little is known about the rates at which de novo transcripts are gained and lost. To detect de novo transcripts that are transiently gained on a short evolutionary time scale, we use DNA long reads and RNA short reads from samples derived from seven Drosophila melanogaster from different geographical locations. Transcripts from the seven samples were distributed into orthogroups according to three newly proposed definitions. Overall, each sample contains between 2,320 and 2,809 unspliced de novo transcripts with most of them being specific to an individual sample. Depending on the definition of transcript orthogroups, we estimate that between 0.13 and 0.34 transcripts are transiently gained per year and that each gained transcript is transiently lost at a rate between 6.6*10^-5 and 2*10^-4 per year. This suggests frequent exploration of new genomic sequences inside species mediated by a high turnover of transcripts. Our study therefore provides novel insight on non-genic transcript dynamics on a very short evolutionary time scale with implications for the process of de novo gene birth.

---

## Enzyme activity versus enzyme stability in variants of laccase CotA from *Bacillus coagulans*

Katrin Gesing*, Sebastian Schreiber, Florian Lenz, Joachim Jose
*University of Münster, Institute of Pharmaceutical and Medicinal Chemistry,*
*PharmaCampus, Correnssstraße 48, 48149 Münster, Germany*
**katrin.gesing@wwu.de*

The laccase CotA from *Bacillus coagulans* was displayed at the surface of *E. coli* and engineered for higher activities at moderate temperatures (30 °C), resulting in the variants CotA T260S and CotA T260S/L385K/F416R. Here, these surface-displayed variants were characterized in terms of their enzyme kinetic constants ($K_M$ and $V_{max}$) and thermal inactivation to gain insights into the molecular basis for the increase in activity.
$V_{max}$ of both variants at 30 °C with 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) as substrate was higher than $V_{max}$ of the ancestor CotA (CotA: 85 mU/mL$_{OD1}$, T260S: 165 mU/mL$_{OD1}$, T260S/L385K/F416R: 228 mU/mL$_{OD1}$). The same was observed for 2,6-dimethoxyphenol (2,6-DMP) as substrate (CotA: 33 mU/mL$_{OD1}$, T260S: 52 mU/mL$_{OD1}$, T260S/L385K/F416R: 383 mU/mL$_{OD1}$). $K_M$ with ABTS was higher than $K_M$ of the ancestor CotA for both variants (CotA: 0.44 mM, T260S: 1.25 mM, T260S/L385K/F416R. 2.08 mM).

With 2,6-DMP, $K_M$ of CotA T260S was slightly higher than for ancestor CotA (CotA: 3.79 mM, T260S: 4.68 mM), whereas $K_M$ of CotA T260S/L385K/F416R was lower (0.31 mM). By the kinetic constants $K_M$ and $v_{max}$, the thermodynamic mechanism of the improved catalysis was analyzed by comparison of the difference in activation energy $\Delta G^{\#}$, total activation energy $\Delta G^{T}$ and change in free energy upon binding ($\Delta G^{B}$) between the reaction catalyzed by ancestor CotA and the variants. With ABTS a destabilization of the ground state appeared to be mainly responsible for the activity increase, whereas with 2,6-DMP a stabilization of the transition state seemed to be the main factor for the increased activity at moderate temperatures. In consequence, the increase in activity was due to two different effects for the two substrates.
Thermal inactivation was measured for incubation at 60 °C. Thermal inactivation of CotA and CotA T260S was nearly identical, whereas CotA T260S/L385K/F416R lost more activity during extended incubation at 60 °C (remaining activity after 16 h: 76 % (CotA), 69 % (T260S), 22 % (T260S/L385K/F416R)). Hence, in case of T260S enzyme activity was increased without reduction of thermostability.

## Goddard - a putative de novo protein important for male Drosophila

Andreas Lange, Prajal Patal, Geoffrey Findlay, Erich Bornberg-Bauer
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Over the past decade, evidence has accumulated that new protein coding genes can emerge de novo from previously non-coding DNA. These proteins are called de novo proteins. Their properties are still poorly understood, and their experimental analysis faces major obstacles. One example of a putative de novo proteins and its successful expression, solved function and partly solved structure is Goddard in *Drosophila melanogaster*. Goddard protein localizes to elongating sperm axonemes and in its absence, elongated spermatids fail to undergo individualization. By combining modelling, NMR and circular dichroism (CD) data, we show that Goddard protein contains a large central $\alpha$-helix, but is otherwise partially disordered. We find similar results for Goddard's orthologs from divergent fly species and their reconstructed ancestral sequences. Accordingly, Goddard's structure appears to have been maintained with only minor changes over millions of years.

## Using Phage display to determine protein-protein interactions of de novo proteins

Andreas Lange*, Priyanka Madhu, Ylva Ivarsson & Erich Bornberg-Bauer
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Comparative evolutionary genomics has revealed that novel protein coding genes can emerge randomly from non-coding DNA. While most of the transcripts which continuously emerge vanish rapidly, some attain regulatory regions, become translated and survive. These so called de novo proteins are small, have little to no structural elements, and their function is still poorly understood. However, proteins of all different kinds of structures or even proteins without defined structure are capable of binding ligands at impressively high levels of both specificity and affinity. Protein-protein interactions (PPIs) are immensely important, driving many chemical reactions and other cellular processes, and are easy to attain compared to enzymatic functions. Since de novo proteins are very young and small genes, PPI is maybe a possible way for them to interact in the organism. Therefore, Proteomic peptide-phage display (ProP-PD) is a fast and easy to use method to measure hundreds of possible binders against putative de novo proteins. ProP-PD is based on classical phage display combined with a combination of computational library design, custom oligonucleotide library synthesis, next-generation sequencing (NGS), and integrative data analysis. Our aim is to identify possible binding motifs for which we could search in the human proteome in order to find possible interaction partners and where our candidate proteins are located in the human body. Hence, we selected ten putative human de novo proteins out of a list of 300 candidates for experimental characterization. The candidate proteins all differ in secondary structure content (α-helix/β-sheet), disorder level, and size. All candidate proteins have been expressed in vivo using E.coli T7 express cells, GroEL chaperone system, and a purification tag (6xHistag or GST-tag). Five out of ten candidate proteins could be expressed with decent amount of concentration (3 mg/mL), and tested with ProP-PD measurements (Uppsala). ProtP-PD or peptides in general have never been used for finding interaction partners or binding motifs for de novo proteins so far.

References:
[1] Bornberg-Bauer, E, Hlouchova, K, Lange A Structure and Function of Naturally Evolved de novo Proteins COSB, 2021,
[2] Ferrando J, Solomon LA. Recent Progress Using De Novo Design to Study Protein Structure, Design and Binding Interactions. Life (Basel). 2021 Mar 10;11(3):225
[3] Benz, C., Ali, M., Krystkowiak, I., Simonetti, L., Sayadi, A. et al. Proteome-scale mapping of binding sites in the unstructured regions of the human proteome Molecular Systems Biology, EMBO. 18(1) 2022.
[4] Eicholt, LA, Aubel, M, Berk, K, Bornberg-Bauer, E, Lange, A. Heterologous expression of naturally evolved putative de novo proteins with chaperones. Protein Science. 2022; 31( 8):e4371.

---

### ***A proteomics approach to immune priming in Tribolium castaneum***

Zoe M Länger, Moritz Baur, Ana Korsa , Jürgen Eirich, Joachim Kurtz
*Institute for Evolution and Biodiversity, University of Münster, Germany*

Invertebrates lack the molecular machinery of the adaptive arm of the vertebrate immune system but were nevertheless shown to be capable of forms of immune memory, which can be induced by prior contact to pathogen-derived cues. However, little is known about the proteins involved in the induction of such phenomena of so-called immune priming. In this work we use the model system *Tribolium castaneum* (*Tc*) and its pathogen *Bacillus thuringiensis tenebrionis* (*Btt*). Larvae of Tc fed with flour mixed with supernatant of the virulent *Btt* strain are better able to survive a subsequent infection with *Btt* spores, in comparison to larvae fed with flour mixed with supernatant of a non-virulent *Bt* strain. To identify the pathogen-derived components needed to induce this immune priming response we used mass-spectrometry based proteomics to compare the proteome of *Btt* with the non-priming inducing strain. Among the potentially immunostimulatory proteins in the *Btt* supernatant, we identified the Cry3Aa toxin, an important plasmid-encoded virulence factor of *Btt*. Upon ingestion this crystal toxin solubilizes and damages the gut epithelium through forming pores thereby contributing to the infection success of *Btt.* Different cry toxins are encoded by different strains of *Bt* and are host-specific in their toxicity, highlighting their potential to induce a specific immune response. In further infection and survival assays we found an association between the presence of the Cry3Aa encoding plasmid of *Bt* and the ability to induce immune priming in *Tc*.

---

## Protein engineering of a surface displayed thermophilic β-glucosidase

Florian Lenz, Paul Zurek, Martina Umlauf, Iasson E.P. Tozakidis, Joachim Jose

*University of Münster, Institute of Pharmaceutical and Medicinal Chemistry, PharmaCampus, Corrensstraße 48, 48149 Münster, Germany*

For the degradation of cellulose e.g. from plant waste, different cellulases are required that need to act in concert. These are endocellulases that randomly cleave internal glycosidic bonds, primarily at amorphous regions of cellulose the polymer, exocellulases that produce disaccharides from the accessible chain ends and finally a β-glucosidase that hydrolyzes the disaccharide to yield glucose. In this process, a low activity of β-glucosidase often represents the limiting factor. This bottleneck was addressed by protein engineering of the thermophilic β-glucosidase BglA from *Caldicellulosiruptor saccharolyticus* (*Cs*BglA) with a temperature optimum of 80 °C towards high activity at 55 °C – the temperature at which cellulose saccharification typically is performed. The variant CsBglA L64R Y73F T221N H324L (*Cs*BglA LYTH) resulted by the merge of beneficial mutations identified in a semi-rational strategy (T221N) and a directed evolution strategy (L64R Y73F H324L). *Cs*BglA LYTH exhibited an increased activity of 552 U/mg (*Cs*BglA 359 U/mg) with cellobiose and an increased activity of 790 U/mg (*Cs*BglA 164 U/mg) with the artificial substrate 4-nitrophenyl β-D-glucopyranoside (*p*NPG) at 55 °C. Its temperature optimum was shifted to 65 °C. $k_{cat}/K_M$ was 2.5-fold increased for cellobiose and 2.6-fold with *p*NPC compared to *Cs*BglA. A residual product inhibition by glucose ($IC_{50}$ = 0.29 M) was determined, making *Cs*BglA LYTH well suited for a combined use with other cellulases in the degradation of cellulose.

---

## Identification and Optimization of a Metagenomic Esterase with Activity towards Polylactic Acid Hidrolysis

Aleksandra Maršavelsk*i, Eva Josić, Marko Močibob, Ivana Kekez, Margaux Aubel, Erich Bornberg-Bauer
**Department of Chemistry, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia*

Polylactic acid (PLA) is a biodegradable polyester with high potential for applications in many fields.[1] The low biodegradability of PLA in natural environments limits its use as a sustainable alternative to conventional plastics. However, microbial enzymes offer a promising strategy to enhance biodegradation of PLA. Accordingly, we applied a metagenomic approach to identify a novel esterase enzyme, capable of cleaving PLA. Search of metagenomic sequence databases identified an esterase, and its activity towards PLA was demonstrated using emulsified PLA in an agarose test.[2] Subsequent crystallographic studies to obtain the enzyme's structure for further optimization resulted in multiple crystal hits. The structure determination process is ongoing, with initial results indicating the enzyme belongs to the α/β-hydrolase family. Ancestral sequence reconstruction was applied to obtain more robust variants stable at higher temperatures. These variants will be ordered as synthetic genes, expressed, and tested for improved thermostability and activity.

[1] Rosenboom, JG., Langer, R. & Traverso, G. Bioplastics for a circular economy. Nat Rev Mater 7, 117–137 (2022). https://doi.org/10.1038/s41578-021-00407-8

[2] Teeraphatpornchai T, Nakajima-Kambe T, Shigeno-Akutsu Y, Nakayama M, Nomura N, Nakahara T, Uchiyama H. Isolation and characterization of a bacterium that degrades various polyester-based biodegradable plastics. Biotechnol Lett 25, 23-8 (2003) https://link.springer.com/article/10.1023/A:1021713711160

---

## Major changes in domain arrangements are associated with the evolution of termite castes

Alina Mikhailova
*Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany*

Domains as functional protein units and their rearrangements along the phylogeny can shed light on the functional changes of proteomes associated with the evolution of complex traits like eusociality. This complex trait is associated with sterile soldiers and workers, and long-lived, highly fecund reproductives. Unlike in Hymenotpera (ants, bees, and wasps), the evolution of eusociality within Blattodea, where termites evolved from within cockroaches, was accompanied by a reduction in proteome size, raising the question of whether functional novelty was achieved with existing rather than novel proteins. To address this, we investigated the role of domain rearrangements during the evolution of termite eusociality. Analysing domain rearrangements in the proteomes of three solitary cockroaches and five eusocial termites, we inferred more than 5000 rearrangements over the phylogeny of Blattodea. The 90 novel domain arrangements that emerged at the origin of termites were enriched for several functions related to longevity, such as protein homeostasis, DNA repair, mitochondrial activity, and nutrient sensing. Many domain rearrangements were related to changes in developmental pathways, important for the emergence of novel castes. Along with the elaboration of social complexity, including permanently sterile workers and larger, foraging colonies, we found 110 further domain arrangements with functions related to protein glycosylation and ion transport. We found an enrichment of caste-biased expression and splicing within rearranged genes, highlighting their importance for the evolution of castes. Furthermore, we found increased levels of DNA methylation among rearranged compared to non-rearranged genes suggesting fundamental differences in their regulation. Our findings indicate an importance of domain rearrangements in the generation of functional novelty necessary for termite eusociality to evolve.

Maximilian Rüttermann, Karthik Suramaniam

*Center for Soft Nanoscience and Institute of Medical Physics and Biophysics*