The noncoding genome, a reservoir of genetic novelty

Christos Papadopoulos

Thèse de doctorat

NNT : 2022UPASL001

université PARIS-SACLAY

Le génome non codant, réservoir de nouveauté génétique

The noncoding genome, a reservoir of genetic novelty

Thèse de doctorat de l'université Paris-Saclay

École doctorale 577, structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Sciences de la vie et de la santé
Unité de recherche : Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Paris-Saclay, le 07/01/2022, par

Christos PAPADOPOULOS

Composition du Jury

Purificación LOPEZ-GARCIA Directrice de recherche, Université Paris-Saclay (ESE)	Présidente
Juan CORTES Directeur de recherche, CNRS (LAAS)	Rapporteur & Examinateur
Joseph SCHACHERER Professeur, Université de Strasbourg (GMGM)	Rapporteur & Examinateur
Jacques CHOMILIER Directeur de recherche, CNRS (IMPMC)	Examinateur
Pierre TUFFERY Directeur de recherche, Université de Paris (MTi)	Examinateur

Direction de la thèse

Olivier LESPINET Professeur, Université Paris-Saclay (I2BC)	Directeur de thèse
Anne LOPES Maître de conférence, Université Paris-Saclay (I2BC)	Co-directrice de thèse

Acknowledgments

Everything started three and a half years ago. I had just returned from a failed interview for a PhD position in Germany and I was really disappointed. But I had to continue searching for a PhD. That was when I met Anne in the office of a colleague, and I told her my story. As by chance she was also searching for a PhD candidate for a new project on the noncoding genome which I found really interesting. The same night I applied for the position and soon we started meeting in cafés all around Paris preparing my presentation for the day of the audition. The day came, everything went well, and I had my PhD scholarship. Three years later, and after a pandemic crisis which changed once and for all our everyday life, here I am writing these lines at the end of the writing of my PhD thesis.

I want to say a big thank you, from the bottom of my heart to Anne who trusted me in this innovative project that was also a big challenge for herself. For all the uncountable hours spent mostly on the phone, due to the pandemic, discussing about philosophical questions around the evolution but also for all the personal advice that she gave to me during these years. She might doesn't know it, but she helped me to evolve as a scientist but also as a personality and I will keep all her words for the rest of my life. Anne is a person passionate about life and science, that can find the beauty of nature hidden in the simplest things around her. One thing I deeply admire at her, is the respect she has for all the results (positive or negative) and that the results which are difficult to interpret instead of scaring her, are the ones that passionate her the most. She never made me feel like a student that has to learn from her wisdom but mostly like a colleague that together we explore the unknown world of the noncoding genome. Every intense phone call that was not arriving to an agreement but mostly to an argument, was forgotten the same moment and the next day we were ready to start again. I am grateful for having met her and for sharing this experience next to her. I know I have a valuable colleague but also a friend for the years to come and she does also.

I would like to thank the Director of the team BIM, Olivier Lespinet for giving me a place in his team during these three years and being always there finding solutions to any problem. All the members of the bioinformatic teams of I2BC (Melina, Leonor, Christine, Claire, Fabrice) for the nice moments that we had together and the funny discussions

during the lunch time. Special thanks to the other PhD students of the team that we shared these three special years: Jean-Noël, Nicolas, Hugo, Paul, Vahiniaina. Also, I need to thank some colleagues from I2BC who I met during these years and truly appreciate: Cécile Fairhead, Fabienne Malagnac, Fabrice Confalonieri, Marie-Hélène Mucchielli. The école doctorale SDSV for giving me the opportunity to work on this innovative project and especially Pierre Capy, Jean-Luc Pernodet and Sandrine Le-Bihan for helping me with all the administrative staff.

Moreover, I must thank the members of my PhD committee who were there every year to follow the advancement of this work and help me with their valuable advice: Cécile Neuvéglise, Daniel Gautheret and Raphaël Guerois. I should not forget all the collaborators who participated in the publication of this work: Isabelle Callebaut for generously offering us HCA and her valuable help in understanding the Hydrophobic Clusters, Jean-Christophe Gelly for the interesting conversations and his expertise on machine learning, Olivier Namy, Isabelle Hatin and Maxime Renard for the experimental part which gave to this work a more integrative point of view.

I would like to thank all the members of the jury for accepting our invitation and participating in the evaluation of this work: Purificación López García, Pierre Tuffery, Jacques Chomilier and especially the two reviewers who spent some of their valuable time to read this manuscript Juan Cortes and Joseph Schacherer.

A big thank you to my family back in Greece for always supporting me and believing in me and especially to my mum for always being there answering to my phone calls. Finally, this journey wouldn't be the same without some special people who made it easier and happier. The biggest thank you I owe it to all my friends in Paris but also abroad for sharing these beautiful years together and for all their patience when I was stressed and anxious and nerve racking!!!! These are Andromachi, Sofia, Isabela, Christina, Panagiotis and Andreas. Last but not least, I would like to thank especially Carlos who has been there, next to me, since the beginning of this experience supporting me and encouraging me in every step.

Table of Contents

Abbreviations

2D	Two Dimensions
3D	Three Dimensions
A	Adenine
AltORF	Alternative Open Reading Frame
ancIGORF	Ancestral Intergenic Open Reading Frame
C	Cytocine
cDNA	complementary DNA
CDS	Coding Sequences
CV	Cross Validation
DIBS	Disordered Binding Site
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA Elements
FN	False Negative
FP	False Positive
G	Gouanine
HCA	Hdrophobic Clusters Analysis
IDP	Intrinsically Disorder Protein
IGORF	Intergenic Open Reading Frame
indels	Insertions-Deletions
lncORF	Long noncoding Open Reading Frame
lncRNA	Long noncoding RNA
MD	Molecular Dynamics
MHC-I	Major histocompatibility complex class I
mRNA	Messenger RNA
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information
ncRNA	Noncoding RNA
NDR	Nucleosome-depleted region
NGS	Next Generation Sequencing
NMD	Nonsense-mediated RNA decay
NMR	Nuclear Magnetic Resonance
nr	Non-redundant
ORF	Open Reading Frame
PCA	Principal Components Analysis
PDB	Protein Data Bank
PPI	Protein-Protein Interaction
ppm	Parts per million
pre-mRNA	premature mRNA
RBF	Radial Basis Function

RNA	Ribonucleic acid
RPF	Ribosome Profiling Footprint
S3	Structure Stable in Solution
SCOP	Structural Classification Of Proteins
SEP	smORF-encoded peptides
P-site	Peptidyl-site
smORF	small Open Reading Frame
snoRNA	small nucleolar RNA
snRNA	Small nuclear RNA
SVM	Support Vector Machine
T	Thymine
TF	Transcription Factor
TM	Transmembrane
TN	True Negative
TP	True Positive
TRG	Taxonomically Restricted Genes
tRNA	transfer RNA
uORF	upstream Open Reading Frame
UTR	Untraslated Region
UV	Ultra Violet

1 Introduction

1.1 What is a gene?

The word gene comes from the greek word *γένεση* (meaning birth) or *γένος* (meaning generation and origin) and was initially coined to denote an abstract "unit of inheritance". The first one ever described these inheritance units, even though he never mentioned the word "gene" but rather the notion of "cell elements", was the so-called "father of genetics" Gregor Mendel in 1865 (Mendel 1865). By the middle of the nineteenth century, scientists observed the cell division and understood that the hereditary information was located in the cell nucleus. However, the physical hereditary material remained unknown. The discovery of the chromosomes by Walther Flemming in 1882 (Flemming 1882) (Flemming actually discovered the chromosome, but the term was proposed a few years later by Heinrich Waldeyer (Waldeyer 1888)) soon provided a fundamental and concrete material for the hereditary factors of Mendel and permitted the proposal of the chromosome theory by the German biologist Boveri and the American geneticist Sutton during the years 1902–1903 (Boveri 1902, 1903; Sutton 1903). Shortly after the birth of the chromosome theory, the phenomenon of gene linkage (Bateson et al. 1909) demonstrated that genes exhibiting "coupling" (and as result were co-transmitted) were located on the same chromosome while genes showing independent assortment were located on different chromosomes.

The term "gene" was coined early in the 20th century, by the Danish botanist Johannsen together with the notions of genotype and phenotype (Johannsen 1909). Thus, by the early 1930s, and thanks to the breakthrough theories of Muller, the concept of the gene became more concrete. Genes were considered as indivisible and dimensionless units of inheritance, each one located at a specific point on a chromosome. They were defined by their four characteristics: (1) hereditary transmission, (2) genetic recombination, (3) mutation, and (4) gene function (Portin and Wilkins 2017). Furthermore, in 1927 Muller associated the concept of the gene with the theory of evolution, while he described the gene as the elementary unit of evolution and the origin of life itself (Muller 1927; Carlson 1966). By the early 1940s, the genetic recombination revealed that the genes could be dissected into segments, converting them from dimensionless points to entities with

length. The ultimate breakthrough for the DNA theory of inheritance was the revelation of the double-helical structure of DNA by Watson and Crick in 1953 (Watson and Crick 1953a, 1953b). The earlier hypothesis "one gene-one enzyme" proposed by Beadle and Tatum in 1941 and which highlighted the connection between genes and proteins, was now proven by the genes' transcription to mRNA (Beadle and Tatum 1941). Soon, the discovery of the universal genetic code by several teams, revealed that the nucleotide sequences are the ones that determine the sequence of polypeptide chains. By the early 1960s, the gene had acquired a definitive molecular identity as a discrete sequence on the genomic DNA molecule that codes for a functional polypeptide product (Gerstein et al. 2007). In the early 1970s the revolution of molecular biology led to the sequencing of the first gene and later of the first genome from the bacteriophage MS2 (Fiers et al. 1971, 1976). In parallel, computational tools were developed for the identification of genes based on their sequence characteristics permitting the detection of the genes by their predicted sequences rather than by their outcome phenotype to the organism (Gerstein et al. 2007). Soon, the advancements in DNA sequencing technologies and later the powerful next-generation sequencing methods led to the explosion of genome projects and consequently to the sequencing of multiple genomes (Hu et al. 2011). The availability of multiple sequenced genomes advanced the field of comparative genomics permitting the identification of most genes by their similarity to other known genes (Gerstein et al. 2007).

However, the hypothesis of "one gene - one mRNA - one polypeptide" started to expire with the advance of high throughput "omics" methods (such as transcriptomics, translatomics, proteomics). In 1986 it was shown that in many cases, a single gene could produce more than one mRNAs through the alternative splicing procedure (Leff et al. 1986). In fact, the genes of eukaryotic organisms are not continuous Open Reading Frames (ORFs) but rather are interrupted by nucleic sequences called introns. Split genes are transcribed into one pre-mRNA molecule, whose introns are removed during the maturation of the mRNA by pre-mRNA splicing. In the alternative splicing, multiple non-consecutive exons are joined together in order to produce the final mature mRNA molecule (Leff et al. 1986; Black 2003). As a result, individual exons can be combined in different ways and produce multiple different mRNAs which are translated into completely different proteins. Additionally to the alternative splicing, the phenomenon

of RNA editing which describes the post-transcriptional molecular processes that modify the structure of the mature mRNA molecule, complicates even more the traditional notion of the gene. In fact, the RNA editing in mRNAs leads into altered amino acid sequence of the encoded protein, different from the one expected by the genomic DNA sequence (Brennicke et al. 1999). Recently, thousands of putative intergenic open reading frames in various eukaryotic organisms have been identified (Hanada et al. 2007; Heinen et al. 2009; Yang et al. 2011; Carvunis et al. 2012; Zhao et al. 2014). Moreover, transcriptomics and ribosome profiling experiments report a widespread transcription of these noncoding regions as well as a pervasive translation of their corresponding RNAs (Kapranov et al. 2002; Clark et al. 2011; Ingolia et al. 2011; Chew et al. 2013; Bazzini et al. 2014). Interestingly, their sequences are more conserved than those of other non-genic sequences suggesting that they could have a functional role (Slavoff et al. 2013). It becomes clear that the classical view of a gene as a unit of heredity aligned along a chromosome, has evolved greatly during the years. Contrary to the Human Genome Project which revealed an impressively low number of protein-coding genes in the human genome, the ENCODE project (ENCODE Project Consortium 2007) highlighted an important number and complexity of the RNA transcripts that the human genome produces, changing dramatically our view of "what is a gene" (Gerstein et al. 2007).

1.2 Small ORFs were systematically being ignored

Efforts to detect and annotate protein coding sequences (called CDS) in genomes using bioinformatic approaches have traditionally relied on arbitrary rules such as amino acid conservation and homology, translation initiation from an AUG start codon and minimum length of 50 or 100 amino acids (Basrai et al. 1997; Couso and Patraquim 2017; Chen et al. 2020). These annotation rules have been widely adopted for convenience and in order to ensure a low number of false positives. They were principally based on the assumption that short peptides are unlikely to fold into stable structures capable of being functional (Couso and Patraquim 2017; Ruiz-Orera and Albà 2019b; Chen et al. 2020). As a result, multiple small ORFs (smORFs), lacking experimental evidence of function, were being systematically discarded and many small proteins remain unannotated (Basrai et al. 1997; Ruiz-Orera and Albà 2019b). Nevertheless, the detection of these smORFs is a complicated task because true conservation and homology of small ORFs is difficult to be

detected due to the tendency of short sequences to present lower conservation scores compared to longer canonical ones (Couso and Patraquim 2017). In addition, many studies in different organisms give evidence for multiple expressed smORFs that do not necessarily initiate from AUG codon (Ingolia et al. 2011; Jackson et al. 2018; Ruiz-Orera and Albà 2019b). As a result, for many years all these intergenic smORFs were lacking annotation and were considered as noncoding sequences. In the 1960s the term "junk DNA" became quite popular while in 1972 the term officially coined independently by Susumu Ohno (Ohno 1972) and David Comings (Comings 1972). This "provocative term" was used to emphasize the "uselessness" of this DNA fraction and for many years the believed dogma was:

Noncoding DNA = Nonfunctional DNA = "Junk DNA"

Nevertheless, millions of smORFs are found in eukaryotic genomes, with thousands of them being mapped to transcripts and some of them fulfilling important physiological functions. This reveals the important transcriptional potential of the genomes which is beyond the already known genes (Couso and Patraquim 2017). Couso and Patraquim (2017) mention that: **"It is as if we have a genome within our genome: a hidden genome about which we know very little"**.

1.3 The noncoding genome is not as silent as believed

1.3.1 Evidence of pervasive transcription

Current estimates indicate that less than 2% of the mammalian genome codes for amino acids in proteins (Clark et al. 2011; Lybecker et al. 2014). However, global transcription profiling as well as mRNA abundance have revealed that the vast majority of the genome is largely transcribed beyond the boundaries of known genes (Kapranov et al. 2002; Clark et al. 2011). This phenomenon is defined as pervasive transcription and is responsible for the generation of a large ensemble of different RNA molecules distinct from those that encode canonical proteins and the ones with already established functions such as tRNAs, rRNAs, snRNAs, and snoRNAs (Jensen et al. 2013).

The first evidence of unconventional transcripts in *S. cerevisiae* was in 2003 through the inactivation of two chromatin remodeling factors (Spt6p and Spt16p). These factors are essential for the correct re-organization of the nucleosomes behind the elongating RNA polymerase II during the transcription procedure. Their inactivation led to spurious intragenic transcription initiated by cryptic promoters within gene bodies(Kaplan et al. 2003). Later, the pervasive transcription was, again, demonstrated in *S. cerevisiae* strains with inactivated certain RNA-degradation pathways. The transcriptome analysis of these strains revealed an ensemble of "hidden transcripts" which normally do not reach detectable concentration levels in wild-type cells (Davis and Ares 2006).

Strikingly, most of the pervasive transcripts in *S. cerevisiae* seem to result from divergent transcription from gene promoters, supporting that gene promoters have an intrinsic bidirectional character and that their apparent directionality is mostly the result of the instability of one of the divergent transcripts (Neil et al. 2009; Xu et al. 2009; Jensen et al. 2013). This divergent transcription originates mostly from nucleosome-depleted regions (NDRs) at the 5' and 3' ends of genes (Jensen et al. 2013). Especially, in the case of dense genomes (such as *S. cerevisiae*), an NDR at the 5' of a gene, would be at the same time an NDR at the 3' of its upstream gene (if the genes have the same sense). The divergent transcription downstream the gene's promoter will have as result the pervasive transcription (from the 3' end) of the preceding gene together with the whole intergenic region of the two genes (Jensen et al. 2013).

In 2012, the ENCODE project (ENCODE Project Consortium 2007) reported that 76% of the human genome's noncoding DNA sequences were transcribed and that almost half of the genome was accessible to transcription factors (Gerstein et al. 2007). cDNA analyses in mouse from different tissues and developmental stages have revealed that at least 63% of the genome is transcribed (Okazaki et al. 2002). Interestingly, thousands of novel protein-coding transcripts were identified as well as around 30.000 long noncoding intronic and intergenic transcripts named long noncoding RNAs (lncRNAs) with no clear protein-coding potential (Carninci et al. 2005; Guttman et al. 2010; Ingolia et al. 2014). Various studies on lncRNAs have proven their multiple roles in cellular functions. LncRNAs have been found to regulate chromosome architecture, to modulate chromosomal interactions, to regulate the recruitment of chromatin modifiers, to act as

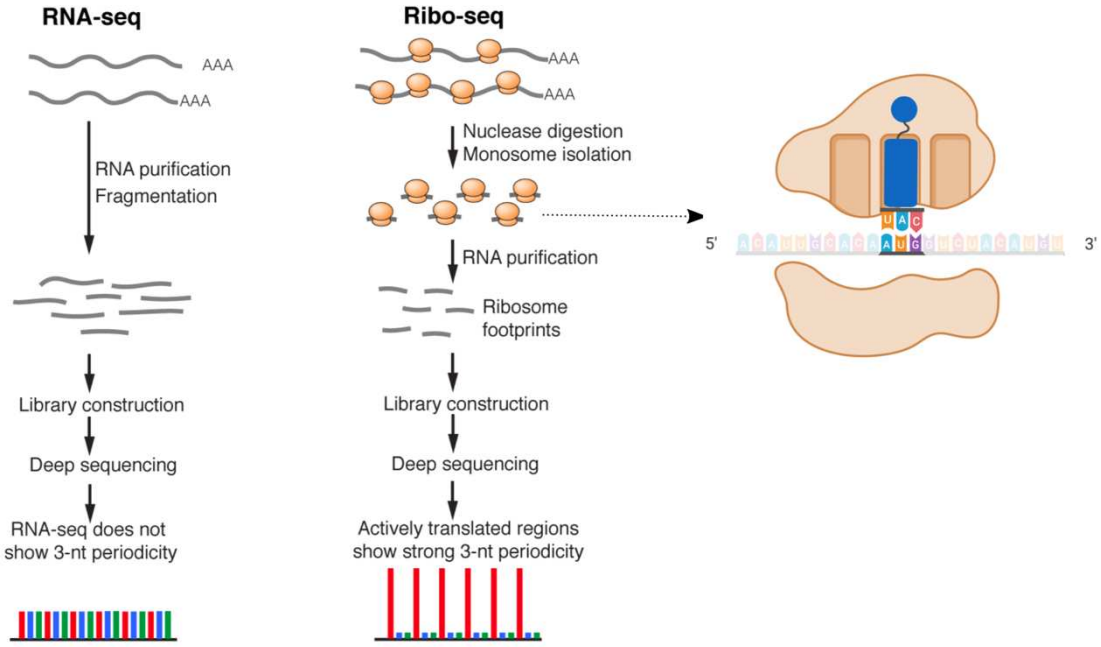
architectural RNAs or even to regulate the translation procedure in the cytoplasm (Yao et al. 2019). Pervasive transcripts in bacteria could also function as sponges for proteins or other ncRNAs (Lybecker et al. 2014). A recent study revealed differential expression of multiple lncRNAs in different human organs and developmental stages, enforcing their important role to the mammalian development (Sarropoulos et al. 2019).

All these observations, of transcription beyond the protein coding genes, gave an important regulative role to the noncoding genome. It became clear that the, so far, called "Junk DNA" was not as useless and inactive as previously thought and that it could participate in multiple cellular functions. Lybecker et al. (2014) proposed that additionally to their regulatory role, some of these transcripts could potentially code for small peptides, further increasing the protein-coding potential of the genomes.

1.3.2 Evidence of pervasive translation

The translatome of an organism or a cell is defined as the ensemble of RNA sequences which are translated by the ribosomal machinery (Ruiz-Orera and Albà 2019b). The explosion of translatomics occurred thanks to a recent and powerful technique called Ribosome Profiling (or simpler RiboSeq) which provides genome-wide snapshots of translation (Ingolia et al. 2009, 2011; Chew et al. 2013; Aspden et al. 2014; Bazzini et al. 2014; Ruiz-Orera and Albà 2019b). Contrary to RNA sequencing (RNA-seq), which aims at targeting complete RNA sequences, ribosome profiling is a very sensitive method, which targets specifically ribosome-protected RNA fragments (Ingolia et al. 2009, 2014; Ruiz-Orera and Albà 2019b; Blevins et al. 2019). In fact, when ribosomes bind a mRNA, they can protect mRNA fragments of around 30 nucleotides from RNAse degradation (François et al. 2021). Rapid translation inhibition through flash freezing, permits to capture a snapshot of ribosome distributions in a particular physiological state of the cell (Brar and Weissman 2015). Nuclease treatment permits the isolation of mRNA fragments (called Ribosome Footprints) corresponding to mRNA regions protected by the ribosome (Brar and Weissman 2015). These fragments are then sequenced and mapped on the reference genome giving information about the translation state of the cell (François et al. 2021). What makes ribosome profiling such a sensitive method is the fact that it has a single nucleotide resolution as it can indicate the precise location of the peptidyl-site (P-

site) of each sequencing read (Ruiz-Orera and Albà 2019b). This is very important because the P-site of the ribosome is indicative of the exact codon which was under translation. Consequently, ribosome profiling not only identifies mRNAs under translation but also specifies the exact frame of the mRNA (among the 3 possible) which was translated and as a result specifies the produced amino acid sequence (Figure 1.1). In addition, ribosome profiling provides an important amount of qualitative information such as translation initiation site, pausing sites, new reading frames, stop codon read through or ribosome residence time (François et al. 2021).



Adapted from Hsu *et al.* 2016

Figure 1.1. Schematical representation and comparison of the RNA-seq (left) and the Ribo-seq (middle) protocols. For the RNA-seq, after RNA purification and fragmentation, the RNA fragments generated are sequenced and mapped on the reference genome. The RNA fragments do not contain any information about the frame of translation and consequently no 3-nucleotides periodicity can be observed. Concerning the Ribo-seq, the RNA fragments (called ribosome footprints) are purified after nuclease digestion and monosome isolation. The RNA fragments purified, are the ones "protected" by the ribosomal machinery and as a result correspond to mRNAs that were under active translation. After their sequencing, the ribosomal P-site on the RNA fragments can be detected through simple position reduction (right). Then, the RNA fragments are mapped on the reference genome but this time with a single-nucleotide precision (P-site). As a result, we do not simply map RNA reads but the exact codon which was under translation and

consequently 3-nucleotides periodicity can be observed on the mapping. The figure of the protocols was adapted from the study Hsu et al. (2016) with the title "*Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis*" and the figure of the ribosome P-site detection was created with BioRender.com.

Ingolia et al. (2009) was the first study that introduced the ribosome profiling as a technique for monitoring protein translation, thus proposing the first in-depth analysis of the translation process in the yeast. Their results highlighted that 75% of the ribosome-protected fragments started on the first nucleotide of a CDS codon (indicating the frame of translation) and revealed a widespread non-AUG translation initiation under starvation stress conditions. The last few years, the translatome of many eukaryotic organisms has extensively been explored and has proven that additionally to sequences encoding classical long proteins (annotated protein-coding genes), the existence of many small ORFs that arrive to be translated, leading to the production of small peptides from presumed noncoding genomic regions (Ingolia et al. 2009, 2011; Ruiz-Orera et al. 2018; Ruiz-Orera and Albà 2019b). These smORFs with translation signatures are principally detected in transcripts previously considered as noncoding. After their identification, many of them have been proven to have important regulatory roles for genes expression (i.e., upstream ORFs which control the expression of other protein coding ORFs), to produce functional micropeptides or even to be simply the result of the pervasive translation of likely nonfunctional proteins (Ruiz-Orera and Albà 2019b). As a matter of fact, an important fraction of translated species-specific smORFs in mouse have been proven to evolve neutrally and consequently no evident functional role can be attributed to their peptides (Ruiz-Orera et al. 2018) while their production can probably be attributed to the pervasive nature of the translation (Ingolia et al. 2014). In fact, the pervasive translation of nonfunctional new peptides can be considered as a waste of energy and material for the cell. However, this less controlled translation in combination with the already mentioned pervasive transcription offers a simple way to the organism for "exploring" the coding potential of its whole genome. Moreover, Ruiz-Orera and Albà (2019) support that species-specific transcripts do not present high levels of expression, something that reduces the cost of their production.

Studies on the transcriptome and the translatome of the yeast have demonstrated the existence of an important number of previously unannotated RNAs with more than 50% of them hosting translated smORFs (Smith et al. 2014). These unannotated RNAs presented an important sensitivity to the nonsense-mediated RNA decay (NMD) pathway, a surveillance mechanism existing in all eukaryotes acting to eliminate mRNA transcripts that contain premature stop codons. Premature stop codons are problematic as they can lead to the production of nonfunctional proteins (Kurosaki et al. 2019). As a matter of fact, the NMD of the unannotated RNAs downregulates the expression of their smORFs in normal conditions while NMD inactivation (or simply dysfunction) could allow small peptides to be produced in the cell (Smith et al. 2014). Notably, Smith et al. (2014) identified in *S. cerevisiae*, 192 unannotated RNAs targeted to rapid decay by NMD in wild-type strains and translated in *upf1Δ* strains where NMD is inactivated. Interestingly, NMD is downregulated under cellular stress conditions or in specific tissues (*e.g.* brain and testes) (Zetoune et al. 2008; Gardner 2008) leading at nongenic ORFs to be translated and produce novel peptides, thus providing additional functional plasticity to the cell.

1.3.3 Evidence of peptides encoded by presumed noncoding regions

The products encoded by smORFs are called smORF-encoded peptides (SEPs) or micropeptides. Mass spectrometry (MS) is the gold standard method for direct detection and characterization of peptides or proteins, even though only few micropeptides have been directly identified by MS (Yin et al. 2019). MS detects smORF-encoded products by matching experimental spectra against theoretical spectra of all candidate peptides represented in a reference or custom database. Potential issues of the method include: sample preparation, low SEPs abundance, small size, short life, usage of alternative start codons (non-AUG) or even tissue/cell/developmental-specific expression patterns (Yin et al. 2019). Ma et al. (2014) highlighted that the low abundance of the SEPs in addition to the stochastic character of the shotgun peptidomics result in the low SEP overlap among samples and different workflows of SEPs detection. In fact, the best strategy for detecting SEPs is the integration of multiple approaches and running multiple replicates (technical/biological).

A proteomic experiment on human cells, identified about 90 new peptides deriving from noncoding regions and which could not be detected into peptides' databases (Slavoff et al. 2013). The size of these peptides ranged between 18 and 149 amino acids, with the majority of them (~80%) being less than 100 amino acids. The abundance of these non-canonical translation products are comparable to those of typical cellular proteins and their sequences are more conserved than those of introns suggesting that they have a functional role (Slavoff et al. 2013). In addition, Ma et al. (2014) identified 237 additional human SEPs by combining different SEP-detection approaches. By analyzing additional cell lines and a tumor biopsy, they observed that in general SEPs are ubiquitous while some of them were specific to a cell line (Ma et al. 2014). Prabakaran et al. (2014) reported through MS experiments 250 non-canonical translation products in mouse neurons coming from both intragenic and extragenic regions. None of these identified peptides showed a similarity to known coding genes products nor to already described peptides in previous MS experiments. One should notice that all of these MS analyses are generally not saturating suggesting that the number of small peptides identified in these studies is largely underestimated (Prabakaran et al. 2014).

For years, the smORFs and their corresponding micropeptides were overlooked due to their small size and the difficulty of their detection (Makarewich and Olson 2017). However, grace to the protemic studies, several micropeptides have been characterized and consequently, the field of peptidomics has attracted more attention. Despite their small size, they have been shown to play critical roles in many biological processes including development, DNA repair, RNA decapping, calcium homeostasis, metabolism, stress signaling, myoblast fusion and cell death (Makarewich and Olson 2017). Some micropeptides, encoded by smORFs within the 5' UTR of genes, often play regulatory roles in gene expression. Similarly, smORF-encoded peptides found within lncRNAs, or overlapping coding regions within mRNAs (alternative ORFs), often function as regulators of transcript stability by engaging the NMD pathway (Yin et al. 2019). In addition, many characterized micropeptides have been shown to bind and modulate larger cellular proteins acting as ligands to receptors or stabilizers of protein-protein interactions (Magny Emile G. et al. 2013; Anderson et al. 2015; Nelson et al. 2016; Makarewich and Olson 2017). Recently, micropeptides (specifically called neoantigens) have been attributed with another crucial role concerning the tumor immunotherapy.

Neoantigens are tumor-specific peptides which are only expressed in tumor cells (Zhang et al. 2021). They can be presented on the cell surface by major histocompatibility complex class I (MHC I) molecules and subsequently recognized by T cells, thus stimulating strong anti-tumor immune response (Makarewich and Olson 2017; Zhang et al. 2021). High-throughput sequencing techniques and MS-based studies enable the screening of smORFs for different species. The several SEPs already identified and characterized indicate that there is an undisclosed world of peptides waiting to be explored. However, how to describe the role and function mode of a validated micropeptide is another big challenge (Yin et al. 2019).

Table 1.1 Advantages and disadvantages of the methods used for the detection of SEPs. The table was extracted and adapted from the study of Yin et al (2019) with the title "*Mining for missed sORF-encoded peptides*"

Method	Advantages	Disadvantages
RNA Sequencing	• Provide a profile of the transcriptome and enable the construction a database reflecting the native transcript composition, including novel sequences. • Allow the detection of peptides containing SNPs associated with diseases. • Enable proteomics studies on non-model organisms with limited genome annotation. • The basis of ribosome profiling and mass spectrometry.	• Laborious, time and money consuming. • Cannot identify alternative start codon sORFs. • Need computational methods to evaluate the coding potential of sORFs. • Need experimental methods to confirm the products of sORFs.
Ribosome Profiling	• Enable direct detection of sORF. • Enable the detection non-AUG sORFs. • Detect 5'-UTR sORFs and 3'-UTR sORFs. • Survey elongation speed, co-translational processing and in organelle protein synthesis.	• Other proteins that bind RNA can cause false positive. • Sample preparation can dramatically impact results. • rRNA and tRNA may cause ribosome profiling noise. • Need computational methods to evaluate the coding potential of sORFs. • Need experimental methods to confirm the products of sORFs
Mass spectrometry	• Detect SEPs directly • Less labor and time consuming	• Sample preparation can dramatically impact results. • Difficult to detect low abundance, small size, short half-life, or tissue/cell- and time-specific proteins.

In a recent work from Cuevas et al. (2021), the authors combined RNA sequencing, ribosome profiling and mass spectrometry in order to detect non-canonical translation products present in whole-cell extracts (proteome) as well as the major histocompatibility complex class I (MHC-I) immunopeptidome of the human. The MHC-I molecules can bind non-covalently various peptides, many of which derive from defective ribosomal products. The MHC-I-peptide complexes are transported to the surface of the cell and can be recognized by the T lymphocytes, a procedure that is called immunosurveillance. As a matter of fact, MHC-I serves as a stabilizing element which favorizes the detection of these noncanonical peptides whose translation product would otherwise be invisible to MS due to their instability and rapid degradation in the cytosol. They identified 1842 new cryptic proteins, 83% of which derived from noncoding ORFs and 17% from the translation of an alternative frame of protein coding ORFs. These cryptic transcripts presented slightly lower abundance and similar translation efficiency with the canonical transcripts. They were shorter and frequently initiated with non-AUG near-cognate codons. Most of the cryptic proteins were only detected in the immunopeptidome of the cells but not in the cytosolic proteome, supporting that they are rapidly degraded due to their unstable character. Interestingly, they observed that cryptic proteins detected in the cell proteome were on average longer than those found in the immunopeptidome supporting that longer peptides can achieve a more stable structure than short ones (Cuevas et al. 2021).

1.4 The noncoding genome contains different types of smORFs

Grace to the advance of multi-omics approaches (transcriptomics, translatomics, proteomics, peptidomics) multiple already presented studies report evidence of widespread transcription of noncoding regions, as well as pervasive translation of their corresponding RNAs, together with evidence of peptides encoded by presumed noncoding regions. These observations support that the noncoding genome hosts an important potential of novel RNA and protein products. Nevertheless, it must be mentioned that, until now the noncoding genome has mostly been studied with an RNA regulatory perspective. However, RiboSeq and MS results have revealed an important potential of smORFs harboring in the noncoding genome highlighting the importance of adopting a more ORF-centered point of view. Hosting a large diversity of ORFs

(regulatory or peptides coding ones) the so-called noncoding genome had to be further characterized. Notably, Couso and Patraquim (2017) combined ribosome profiling, peptide tagging and bioinformatic analyses and proposed the existence of at least five types of smORFs in *Drosophila melanogaster* genome. These different smORFs present distinct characteristics such as transcript organization, size, conservation, mode of translation, amino acid usage and peptide structure properties.

- **Intergenic ORFs:** They are small genomic sequences which occur mostly by random nucleotide permutation in regulatory or "junk" DNA. They do not present transcription or translation signals (in at least detectable levels) and that is why they are mostly considered as non-functional. They constitute the most abundant category corresponding to 96% of the total smORFs (Couso and Patraquim 2017).

- **Upstream ORFs (uORFs):** They are smORFs located in the 5' untranslated regions (UTRs) of canonical protein coding genes. They are thought to have a regulating role by repressing the translation of their downstream ORF. As a result, uORFs are often depleted from regions in proximity to highly expressed coding ORFs. Their repressive effect may occur by several mechanisms such as ribosome stalling, inhibition of translation re-initiation, or uORF induced nonsense-mediated decay (Couso and Patraquim 2017). Even though some have been shown to produce detectable polypeptides (Slavoff et al. 2013), in general uORFs present low conservation levels and their amino acid usage is different from the one of canonical proteins. Consequently, only a small subset of the uORFs encoded peptides are expected to present any function (Ruiz-Orera and Albà 2019b).

- **Long noncoding ORFs (lncORFs):** They are smORFs embedded in putative long noncoding RNAs (lncRNAs). Recently, it has been found that many lncRNAs are likely to encode small proteins or peptides, however a considerable debate remains about whether this translation is productive. Many RNAs initially named as lncRNAs were later shown to translate peptides associated to organisms' development and physiology (Couso and Patraquim 2017). Even though lncRNAs do not present high conservation levels, recently Ruiz-Orera and Albà (2019a) identified 289 mouse lncRNAs which shared homology with human transcripts

indicating that they are likely to be functionally relevant. These transcripts were enriched in translated lncORFs supporting that they are likely to encode small proteins.

- **Short coding sequences (short CDS):** They are short protein coding ORFs with transcripts presenting mRNA characteristics and translation efficiency similar to canonical proteins. Even though hundreds of them exist in vertebrate genomes only a small portion has been functionally characterized (Couso and Patraquim 2017).

- **Short isoforms:** They correspond to alternative transcripts or splice forms of longer, canonical protein-coding ORFs. Their identification is quite difficult because it relies exclusively on experimental data from proteomics and ribosome profiling studies (Couso and Patraquim 2017).

1.5 Pervasive expression of smORFs and genetic novelty

All the previously mentioned studies, give evidence that a non-negligible number of loci outside of the well-defined protein coding regions in bacteria (Ndah et al. 2017; Weaver Jeremy et al. 2019), fungi (Ingolia et al. 2009; Wilson and Masel 2011; Carvunis et al. 2012), animals (Ingolia et al. 2011; Chew et al. 2013; Bazzini et al. 2014; Ingolia et al. 2014; Aspden et al. 2014; Ruiz-Orera et al. 2018) and plants (Hanada et al. 2007; Hsu et al. 2016), are transcribed and translated in a pervasive way, leading to the production of numerous small proteins in the cell. Interesting questions concerning the fate of these small proteins and their impact on the cell can be posed. Indeed, even though it is now clear that the noncoding genome can produce a large number of peptides, among which some of them have been shown to be functional, the fraction of functional peptides among all the pervasively translated products remains unknown. Also, the evolutionary longevity of functional peptides is to be further investigated, along with their contribution in genome and proteome evolution. Indeed, functional peptides have been mostly associated to functions related to adaptation, stress response, signal transduction etc. (Hemm et al. 2008; Fozo et al. 2008; Storz et al. 2014; Orr et al. 2020). One can ask

whether these peptides will be short-lived in evolution or whether they will be fixed and established as novel genes. Precisely, comparative genomics studies over the last few years have revealed multiple examples of functional protein-coding genes with no homologs in other species which have emerged from previously noncoding regions, called de novo genes (Levine et al. 2006; Cai et al. 2008; Li et al. 2009b; Knowles and McLysaght 2009; Li et al. 2010; Murphy and McLysaght 2012; Gubala et al. 2017; Vakirlis et al. 2018; Zhang et al. 2019a).

All these studies show that the so-called noncoding genome is an important reservoir of small ORFs, which upon pervasive transcription and translation can produce an important number of small peptides in the cell's cytosol. Nevertheless, most of these peptides, if not deleterious for the cell, are expected to be short-lived and instantly degraded with not particular functionality. However, numerous examples show that sometimes functional novel products can emerge from this procedure. All these results attribute a central role to the noncoding genome in the emergence of genetic novelty, which upon pervasive translation offers the raw material for selection.

1.6 De novo genes

Nowadays, the increasing number of sequenced genomes in combination with the rapid progress of bioinformatic methods for sequence comparison have led to the expansion of taxonomic sampling and therefore to the significant advance of comparative genomics (Schlötterer 2015). Multiple studies have revealed the existence of numerous genes lacking homologs in any other lineage (orphan genes) or being present only in closely related species (taxonomically restricted genes -TRGs) (Tautz and Domazet-Lošo 2011; Schlötterer 2015). These genes are thought to be particularly important for taxon-specific developmental adaptations and interactions with the environment (Tautz and Domazet-Lošo 2011; Palmieri et al. 2014). Even though up to one-third of the total genomes' genes are TRGs or strictly orphan, their evolutionary origins are still not clear (Tautz and Domazet-Lošo 2011; Palmieri et al. 2014). Tautz and Domazet-Lošo (2011) support that orphan genes could emerge in a genome through two distinct ways: (i) Gene duplication followed by fast divergence to a point that the homology detection tools are

not capable of capturing the similarity signal and (ii) de novo evolution from previously noncoding regions.

In the first scenario (Figure 1.2A), the orphan gene results from the gene duplication or transposition of an already established gene and the fast adaptive evolution of this gene copy until the complete loss of similarity with its parent sequence. Alternative versions of this scenario such as transposon insertions in the protein coding ORF or "overprinting" (expression of alternative ORFs that overlap pre-existing genes) are also possible and could lead to the production of completely different proteins. In the second scenario (Figure 1.2B), random mutations occurring on the noncoding genomic regions, would form spurious cryptic functional sites (i.e., transcription initiation regions or polyadenylation sites) which could lead to the regulated transcription of an RNA molecule. This RNA could either have a structural role (like numerous lncRNAs) or could eventually acquire a functional ORF capable of coding for a new protein (Tautz and Domazet-Lošo 2011).

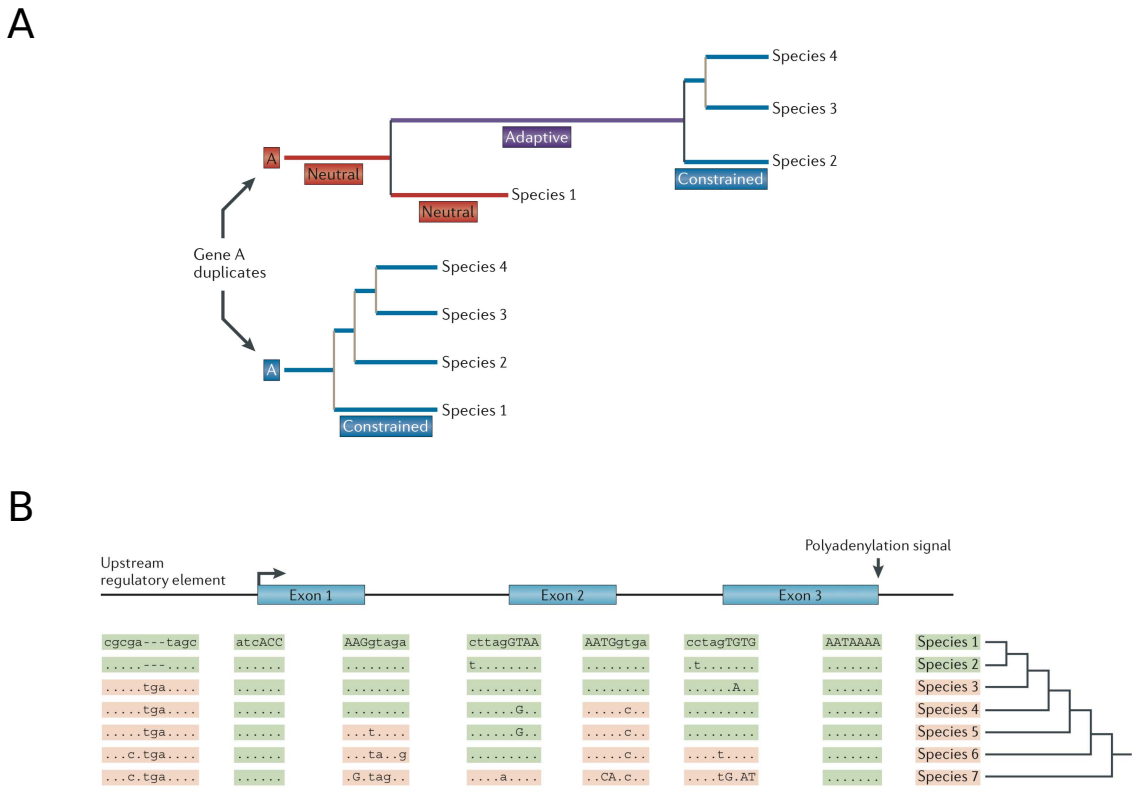


Figure 1.2. (A) Duplication and fast divergence model for orphan gene evolution. Once the gene A is duplicated, the one copy is free to diverge with a neutral rate accumulating random mutations (upper part, red). The freely diverging copy could assume a new function and would go through an adaptive phase in the respective lineage (species 2, 3 and 4). It would diverge to a point that no homology with its parent gene A would be detectable and would thus become an orphan gene. **(B) De novo evolution model for orphan genes.** This example is modelled according to a real case in mice and corresponds to a functional gene that evolved out of a noncoding sequence. Different functional sites of the gene are presented such as an upstream regulatory element before the Exon 1, the transcriptional start site (arrow), the exon junctions (in capitals) and a polyadenylation signal at the end of the Exon 3. The total ORF of the gene is functional only for the species 1 and 2 (contain only green boxes) while for outgroup species (species 3 to 7) at least one element is missing (pink boxes). Both figures were extracted by the study Tautz and Domazet-Lošo (2011) with title *"The evolutionary origin of orphan genes".*

De novo genes arise from DNA sequences that were ancestrally non-genic (Van Oss and Carvunis 2019). Some of their principal characteristics are their shorter ORFs, fewer exons, lower expression levels and similar codon usage compared to CDS sequences (Tautz and Domazet-Lošo 2011; Schlötterer 2015; Van Oss and Carvunis 2019). They are found to evolve more rapidly than established genes and present tissue-specific (brain and testis) or condition-specific (under stress) expression (Levine et al. 2006; Ingolia et al. 2011; Ruiz-Orera and Albà 2019b). For many years, the de novo gene emergence from previously noncoding DNA sequences constituted a rarely observed event and was not considered as a potential evolutionary process of gene birth (Jacob 1977; Siepel 2009; Ohno 2013; Zhang et al. 2019b). Susumu Ohno, in his book "Evolution by gene duplication" (Ohno 2013) supports that all new genes arise from already existing ones while Francois Jacob claimed that "the probability that a functional protein would appear de novo by random association of amino acid is practically zero" (Jacob 1977). However, de novo emergence from noncoding regions has now been proven to be an undeniable additional mechanism and studies reporting evidence of de novo gene birth are published every year, thereby giving a new role to noncoding regions in the creation of genetic novelty (Levine et al. 2006; Cai et al. 2008; Knowles and McLysaght 2009; Murphy and McLysaght 2012; Carvunis et al. 2012; Schlötterer 2015; Gubala et al. 2017; Zhang et al. 2019a; Heames et al. 2020; Bornberg-Bauer et al. 2021). De novo genes have been detected in diverse organisms such as *S. cerevisiae* (Cai et al. 2008; Bungard et al. 2017), *A. thaliana* (Li et al. 2016), *D. melanogaster* (Levine et al. 2006), *M. musculus* (Murphy and

McLysaght 2012) and *H. sapiens* (Knowles and McLysaght 2009). Notably, a pioneering study of Levine et al. (2006) conducted in *D. melanogaster*, identified five novel genes that have derived from ancestrally noncoding sequences. Their experimental results show significant amounts of noncoding DNA transcription though at a low level supporting that noncanonical transcription could occasionally be beneficial, resulting in the "recruitment" of noncoding DNA into novel function and consequently into de novo gene evolution.

1.7 Detection and validation of de novo genes

The numbers of detected de novo genes vary significantly among the different studies due to differences in the search strategies. As a matter of fact, there are significant challenges concerning not only the accurate detection of novel genes but also the validation of their actual de novo emergence from previously noncoding sequences (McLysaght and Guerzoni 2015).

1.7.1 Detection of de novo genes

(i) **Genomic phylostratigraphy**: Genomic phylostratigraphy is a sequence similarity-based method that permits the relative dating of every gene of a given organism. Practically, it involves the detection of all homologous sequences for a given gene (using either sequence-based or more sensitive profile-based detection methods) and the identification of the most distantly related species in which a homolog is detected. Based on a predetermined phylogeny, it becomes possible to assign a relative "age" (or "genomic phylostrata") to every single species of the tree and as a result localize and date the first evidence of existence of every single gene of an organism of interest. In the case that a gene lacks any detectable homolog outside its own genome or closely related species, it can be considered taxonomically restricted or orphan gene, having emerged de novo or not. It becomes evident that the accuracy of this method depends directly on the selection of the species to be compared, the quality of their annotation and their relative evolutionary relationships (Van Oss and Carvunis 2019). Talking specifically about *Saccharomyces cerevisiae*, the studies conducted by Carvunis et al. (2012) and Wu and Knudson (2018) constitute the two major genomic phylostratigraphy analyses that tried

to detect de novo genes by comparing the homology of *S. cerevisiae* ORFs with multiple neighboring yeast species.

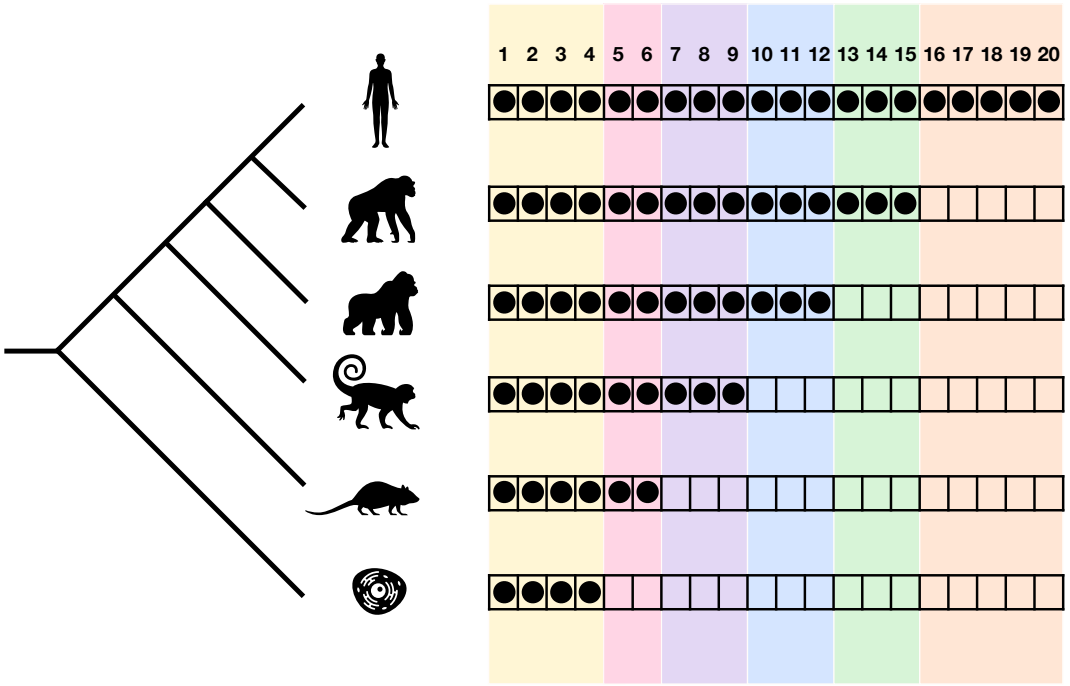


Figure 1.3. Schematical representation of the genomic phylostratigraphy method permitting the relative dating of every gene for a given organism. In this quite simplified example, the human is the reference organism for the relative dating of a subset of its 20 genes. Using homology detection methods, we assign the presence (filled circle) or the absence (empty box) of the 20 genes of interest in the 5 neighboring species selected to be compared. We can observe that there are genes (1 to 4, highlighted in yellow) which are present in all the organisms studied and correspond to ancient genes. The genes 5 and 6 (highlighted in pink) are detected in all organisms except the yeast meaning that they correspond to genes specific to mammalian organisms. The genes 10 to 12 (highlighted in blue) correspond to genes specific to Hominidae etc. It can be observed that the genes 16 to 20 (highlighted in orange) do not present any homolog in any other linage meaning that these genes are Orphan or Taxonomically Restricted Genes of the human.

(ii) **Synteny-based approaches:** These methods permit the identification of the nongenic ancestors of candidate de novo genes through the detection of their homologous noncoding sequences in other species using syntenic sequences. Syntenic sequences are genomic regions in which the order and the relative positions of genomic

elements (i.e., genes, k-mers or exons) which play the role of anchoring "markers" have been maintained during evolution. These methods offer a more accurate way for validating the de novo emergence of a gene and notably differentiate it from an orphan gene by detecting specifically its homologous noncoding region. One major limitation of these methods is the fact that synteny does not stay detectable for long timescales especially in lineages with high rates of chromosomal rearrangements. In addition, it becomes evident that these methods demand high-quality sequenced genomes with fewer fragments (Van Oss and Carvunis 2019). Notably, Vakirlis et al. (2018), Lu et al. (2017) and Vakirlis et al. (2020b) have used syntenic regions alignments in order to detect *S. cerevisiae* de novo genes localized into orthologous intergenic regions.

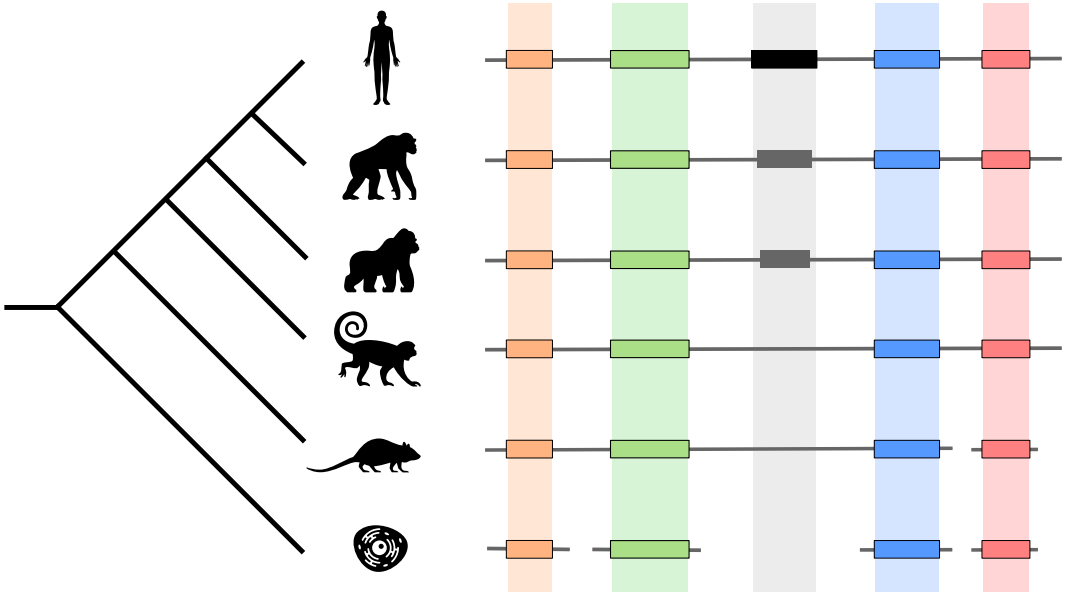


Figure 1.4. Schematical representation of the de novo gene detection with syntenic regions alignment. In this simplified example, the gene of interest is highlighted in black and the other four genes (two upstream and two downstream of the black gene) constitute the anchoring markers. We can observe that grace to the relative positioning of the four anchoring genes we are capable at detecting the orthologous noncoding intergenic region of the black human gene to its closest neighbors (presented in grey) and validating the de novo emergence of this orphan gene. Is important to highlight the subtility of this technique by mentioning that the four anchor genes detected their homologs in the yeast genome but

their relative positioning was lost (synteny rupture) making impossible to detect the orthologous intergenic region of the black gene.

Vakirlis and McLysaght (2019) developed a standardized method using synteny-based approaches in combination with homology searches in an attempt to propose a strategy for more accurate de novo genes identification.

1.7.2 Validation of a de novo gene

Even if the detection of de novo emerging orphan sequences is a difficult task, it is not sufficient to support the existence of a newly evolved gene. That is because the notion of gene is usually related to the notion of function and consequently, every gene is expected to code for a functional product (Van Oss and Carvunis 2019). As a result, in order to validate that the detected de novo sequence is indeed a novel gene, experimental proof of its functional role in the organism is necessary. The first step towards investigating the functionality of the novel sequence, can be the validation of gene expression patterns under normal or stress conditions. Multiple studies have shown that de novo genes present low expression rates which most of the time are detected under specific environmental conditions (Schlötterer 2015). The expression of a novel gene can be verified at multiple levels such as the transcription (with RNA-seq), the translation (with Ribo-seq) or even the detection of a final protein product (with MS or western blotting) (Van Oss and Carvunis 2019). Once the expression of the novel gene is verified, then its functional role in the organism must also be tested, through genetic approaches of overexpression or gene disruption and analysis of their impact to the organism's phenotype or fitness (Kellis et al. 2014; Zhao et al. 2014; Vakirlis et al. 2020a).

On the other hand, evolutionary approaches can also support the functionality of novel genes based on the fact that functional genes are subjected to purifying selection which operates against deleterious mutations in order to maintain their function (Van Oss and Carvunis 2019). As a result, functional genes tend to present lower frequencies of these deleterious mutations as they mostly tend to conserve their protein sequence intact. To do so, the purifying selection favors mostly nucleotide mutations that do not alter the

gene's protein sequence by selecting synonymous (mutations that result to an alternative codon which codes for the same amino acid) rather than non-synonymous (mutations that result to an alternative codon which codes for a different amino acid) codon nucleotide substitutions (Yang and Nielsen 2000). The ratio of the frequencies of these two substitution events (dN/dS) is indicative for the selection type exerted on the gene of question. In the neutral scenario of a sequence which presents no functionality and mutates randomly, we can anticipate similar frequencies of synonymous and non-synonymous substitutions and consequently a dN/dS ratio around the value 1. In contrast, a functional gene under purifying (or negative) selection would present a dN/dS ratio value less than 1 while the opposite is true for genes under positive selection (Yang and Nielsen 2000).

Nevertheless, validating the de novo emergence of an orphan sequence requires not only proving the functionality of the novel gene, but also the lack of functionality for their evolutionary antecedents (Siepel 2009). While determining biological function may be difficult, proving lack of function is even harder (Gerstein et al. 2007). Detection of disabling mutations (absent start codons, premature stop codons or frameshift indels) in the orthologous intergenic sequences of neighboring species could constitute indirect evidence of lack of functionality (Siepel 2009).

1.8 Examples of de novo emerged genes

Together with the above-mentioned work of Levine et al. (2006), Knowles and McLysaght (2009) was one of the first studies that identified de novo emerging genes in the human genome. Starting with a set of 644 human genes absent from the chimpanzee genome and applying various filters proving genes' functionality (i.e., mRNA and protein expression) they identified three candidate human de novo genes. Using syntenic alignments with the chimp and macaque genomes, they detected disabling mutations (indels leading to frame shifts or premature termination) suggesting the lack of functionality in the chimp and macaque orthologs while enforcing the de novo emergence of these three genes (Knowles and McLysaght 2009; Siepel 2009). Two years later, Wu et al. (2011) following a similar protocol but with more enriched gene expression databases, identified 60 de novo human genes. Murphy and McLysaght (2012), following the same method, proposed 69 de novo

candidates in the mouse. However, they were able to identify the orthologous noncoding sequence in rat, guinea pig and human only for 11 of them. For 7 out of the 11 cases, they were able to detect the mutational events that led to birth of the de novo ORF while for the remaining 4 cases the transition from noncoding to ORF was less clear. The same detection protocol combined with a logistic regression model predicting the codability of the novel ORFs were used by Vakirlis et al. (2018) enabling them to propose 30 de novo candidate genes in *Saccharomyces cerevisiae*. Recently, Zhang et al. (2019) identified at least 157 de novo ORFs in *Oryza sativa* with verified recent ancestral noncoding sequences and evidence of translation for 57% of them. Prabh and Rödelsperger (2019) identified 2 de novo genes in *Pristionchus nematodes* with transcription evidence and premature STOP codons in orthologous sequences while Zhou et al. (2008) detected 2 de novo genes in *D. melanogaster* with transcription evidence matching with noncoding regions to neighboring species. Progress in comparative genomics and Next Generation Sequencing (NGS) have enabled the detection of hundreds of de novo genes in multiple studies, thus providing the community with large and well annotated datasets for investigating the mechanism underlying the emergence of de novo genes. Despite all the effort, many of these studies lack evidence for the origination from a noncoding ancestral sequence in their reported cases.

The methods for the detection of de novo genes have become more accurate with the years and have permitted the identification of many novel genes, in different organisms, bearing proof of their noncoding origins. However, the mechanism behind the emergence of de novo genes stays unclear until today. As it has already been discussed, experimental evidence (RNA sequencing and Ribosome Profiling) support that the eukaryotic genomes are largely transcribed and translated in a pervasive way, leading to the expression of numerous unannotated intergenic ORFs (IGORFs) or annotated as lncRNA ORFs (lncORFs). Furthermore, mass spectrometry experiments confirm the existence of many of these translation products in the cell with the identification of hundreds of peptides derived from noncoding regions. All these results attribute a central role to the so-called noncoding genome in the emergence of genetic novelty, which upon pervasive translation offers the raw material for selection. This pervasive expression of intergenic regions has constituted the basis for various theoretical models proposed to describe the de novo birth of genes.

1.9 Models for de novo genes birth

Multiple models, not necessarily exclusive the one to the other, have been proposed trying to explain the mechanism behind the emergence of de novo genes. In his study, Schlöterer (2015) supported that the birth of a de novo gene coding for a functional protein implicates two distinct steps: the establishment of a regulated transcription and the acquisition of an ORF. The order of these two events is not clear, permitting to propose two different but equally possible models. The **ORF first model** stipulates that the emergence of a long de novo ORF precedes the one of its promoter region and that multiple ORFs exist in the genomes "awaiting" the establishment of their regulation (Figure 1.5A). On the other hand, the **transcription first model** relies on the observation that genomes can be pervasively transcribed and stipulates that previously transcribed (and actively translated into short peptides) smORFs, can be subject to selection while acquisition of mutations, such as stop codon mutations, could make them grow into functional de novo genes through combination with neighboring smORFs (figure 1.5B).

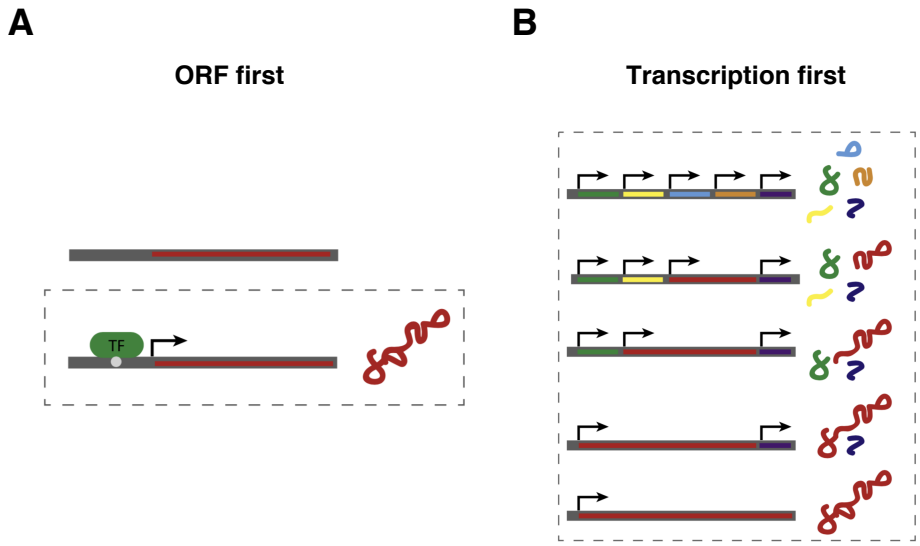


Figure 1.5. (A) ORF first model. A fully functional ORF is present but not expressed due to the lack of regulatory signals. Once a functional transcription factor binding site is generated inside the promoter region, the de novo gene is expressed and translated. **(B) Transcription first model.** Several short peptides are expressed from different smORFs. During evolution and through the acquisition of new mutations the smORFs are combined into a longer protein coding de novo gene. Both figures were extracted by the study Schlöterer (2015) with title *"Genes from scratch – the evolutionary fate of* de novo *genes"*

As a matter of fact, the transcription first model supports that de novo proteins emerge and further elongate through the combination of multiple consecutive smORFs each one capable to code for different short peptides.

This model is in line with another potential mechanism of de novo gene birth (named "grow slow and moult") which is specific to protein-coding genes (Bornberg-Bauer et al. 2015). Based on this model, protein-coding ORFs could eventually expand their ends via occasional and later more constitutive read-through translation leading to the expression of novel N- and C-terminal domains. These novel domains may be well integrated in the preexisting protein structure and/or further be refined by selection potentially offering novel functions to the old protein. Additionally, sequences encoding these novel domains could eventually separate by their hosting ORF leading to the creation of a de novo gene (Bornberg-Bauer et al. 2015; Van Oss and Carvunis 2019). This model is supported by studies conducted on yeast and fly proteins, reporting that the vast majority of proteins' orphan domains are found to be located at the protein termini and therefore suggest that these orphans have been created by mutations affecting the start or stop codons of the preexisting proteins (Ekman and Elofsson 2010; Bitard-Feildel et al. 2015). These observations pose an interesting question about to which extent, these orphan domains could have emerged from noncoding regions, thus attributing a new role to the noncoding genome as a reservoir of novel structural domains which can be grafted on preexisting proteins through termini extension. It reminds an interesting study published in 2015 by Alva et al. reporting the identification of 40 ancestral protein fragments sharing remote homology while occurring in non-homologous domains (Alva et al. 2015). These fragments are widespread in the most ancient folds and may correspond to the vestiges of a primordial RNA-peptide world. Nevertheless, these ancestral peptides presented a wide diversity of amino acid sequences showing that essentially every one of these peptide structures can be formed by a broad range of different sequences. Their results support that the emergence of protein domains, and therefore proteins, might have occur by the repetition, fusion and accretion of these (and other similar) ancestral peptides. A question that arises is to which extend, these peptides extracted by evolutionary established proteins but encoded by a broad range of amino acid sequences could also be hosted by the unevolved noncoding sequences, thus supporting a potential role of the noncoding genome to the structural diversity of the proteins.

In addition to these models, two other theories focusing mostly to the structural properties of the de novo genes have been proposed in order to speculate how a noncoding sequence can be transformed into a protein coding gene. The first one is the "continuum" theory which supports the existence of translated intermediate stages between non-genes and genes, called proto-genes (Carvunis et al. 2012). The later are immature gene-like sequences which can either evolve towards de novo genes or return to their ancestral noncoding state. The second one is the "preadaptation" theory which supports that a novel gene could emerge from a noncoding region only through an "all-or-nothing" transition to functionality (Wilson et al. 2017). This means that de novo genes can occur only from sequences that have been already pre-adapted to not give birth to a harmful product.

In more details, the "continuum" model is based on the hypothesis that genes originated de novo could initially present simple characteristics and gradually become more complex over evolutionary time. The authors introduced the notion of proto-genes which correspond to intermediate and reversible stages of de novo gene birth, mirroring the well-described pseudo-genes stages of gene death. In fact, genes are thought to emerge de novo when non-genic sequences become transcribed, acquire ORFs and the corresponding non-genic transcripts access the translation machinery. Using a phylostratigraphy-based method, Carvunis et al. (2012) classified all the annotated ORFs of *Saccharomyces cerevisiae* into 10 groups based on their conservation (Figure 1.6A). 12% of them were found only in Saccharomyces *sensu stricto* species (mentioned as $ORF_{[1-4]}$) presenting weaker conservation signals thus supporting their recent emergence and their characterization as proto-genes. The remaining 88% of the annotated ORFs (mentioned as $ORF_{[5-10]}$) corresponded to well characterized genes presenting strong conservation signals. In addition to the annotated ORFs, they also extracted all the unannotated intergenic ORFs of more than 30 nucleotides and free from overlap with any annotated feature on the same strand (~108.000 ORFs). These unannotated ORFs (mentioned as $ORF_{[0]}$), were specific only to *Saccharomyces cerevisiae* and corresponded to noncoding ORFs, a subset of which could correspond to initial proto-gene candidates with no conservation. They observed that $ORF_{[1-4]}$ presented intermediate frequencies of amino acids between those of $ORF_{[5-10]}$-encoded proteins and the theoretical translation

products of the $ORF_{[0]}$ category. More precisely, comparison between the $ORF_{[1-4]}$-encoded proteins and $ORF_{[0]}$-encoded peptides did not present any significant difference in frequencies of strong hydrophobic residues (M,I,L,F,W,Y,V). However, this was not the case when they were compared with the $ORF_{[5-10]}$-encoded proteins where the younger $ORF_{[1-4]}$-encoded proteins seem to be enriched in hydrophobic residues than the older yeast proteins (figure 1.6B). This higher hydropathicity of $ORF_{[1-4]}$-encoded proteins was in accordance with their higher tendency to form transmembrane regions and their lower propensity for intrinsic structural disorder in comparison with their older counterparts (Figure 1.6C).

Using ribosome profiling data, they identified 1891 young ORFs (1139 $ORF_{[0]}$ and 752 $ORF_{[1-4]}$) with evidence of translation. Comparing synonymous against non-synonymous mutations of these young translated ORFs, they concluded that the majority of them did not present significant deviation from neutral evolution. However, the fraction of young ORFs under purifying selection was increasing with the conservation level permitting them to propose the continuum model in which young ORFs (proto-genes), upon pervasive translation, can occasionally acquire adaptive functions and then be retained and established as novel genes by natural selection. This continuum was further supported by the identification of transcription factors (TF) suggesting ORFs' regulation. It was found that young proto-genes ($ORF_{[1-4]}$) seem to be regulated by several TFs and that the TFs' number increases with the conservation level, indicating the ORF's integration into larger regulatory networks (Abrusán 2013).

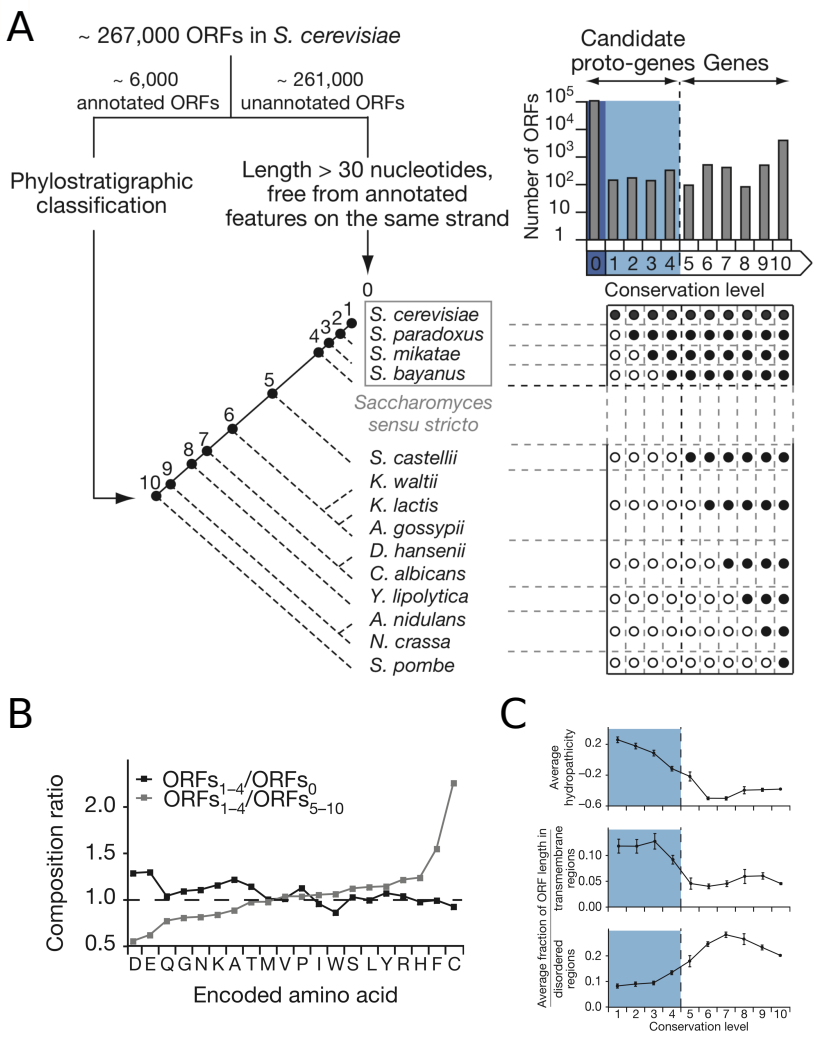


Figure 1.6. (A) Phylostratigraphy protocol used by Carvunis et al. (2012) permitting to assign conservation levels to *S. cerevisiae* ORFs. Conservation levels of annotated ORFs were assigned based on the phylogenetic tree, by inferring their presence (filled circles) or absence (open circles) in the different species. Unannotated intergenic ORFs were assigned to conservation level 0. Unannotated ORFs (level 0) together with young annotated *Saccharomyces sensu stricto* ORFs (levels 1-4) correspond to initial proto-gene candidates. Top right, number of ORFs assigned to each conservation level (logarithmic scale). (B) Amino acids frequency shift with increasing conservation level. For both lines as reference amino acid frequency was taken the one of the $ORFs_{[1-4]}$ and was compared with the frequency of each amino acid in $ORFs_{[0]}$ (black line) and in $ORFs_{[5-10]}$ (grey line). Values higher than 1 indicate amino acids enriched in $ORFs_{[1-4]}$ while values lower than 1 indicate amino acids depleted in $ORFs_{[1-4]}$ and enriched in $ORFs_{[0]}$ or $ORFs_{[5-10]}$, respectively. **(C)** Average hydropathicity (top), average fraction of ORFs predicted as TM (middle) and average fraction of ORFs predicted as disordered (bottom) per conservation level. All the figures were extracted by the study Carvunis et al. (2012) with title "*Proto-genes and de novo gene birth*".

Recently Vakirlis et al. (2020a) proposed the **TM-first** (transmembrane-first) model which constitutes a complementary version of the proto-genes theory. They observed that adaptive incipient proto-genes with transmembrane domains are more likely to be beneficial for the cell (increase fitness) than non-transmembrane ones. They speculate that the membrane environment can provide a safe niche for transmembrane proto-genes, protecting them from proteasome-mediated degradation and preventing deleterious non-specific interactions in the cytoplasm.

Notably, they classified all the annotated sequences of *Saccharomyces cerevisiae* in two groups: (i) de novo **emerging ORFs** identified using a combination of phylostratigraphy and syntenic alignments and (ii) **established ORFs** coding for useful protein products**.** The emerging ORFs did not present evidence of canonical protein-coding genes while their disruption seemed to be inconsequential for the survival of yeast cells. They conducted an overexpression screening analysis in order to detect adaptive emerging ORFs that increased relative fitness upon increased expression. Testing five different environments of varying nitrogen and carbon composition they identified 28 adaptive emerging ORFs that increased relative fitness in at least one environment. These 28 adaptive ORFs presented high thymine content accompanied by high propensity to form transmembrane domains.

In fact, the TM-model constitutes a more specific version of the proto-genes model which supports that thymine-rich intergenic sequences are capable of generating a diverse reservoir of novel peptides with high transmembrane domains propensity. Transmembrane emerging peptides were shown to be beneficial for the cell as they increased the relative fitness of the organism. Upon acquisition of translation and under the effect of adaptive change, these transmembrane peptides could evolve towards more genuine transmembrane proteins. However, this model is mostly based on overexpression screening results, and it should be mentioned that high expression levels of adaptive peptides may never occur in canonical growth or even in natural environment conditions.

On the other hand, contrary to the proto-gene model (and consequently the TM-first model), the **preadaptation model** supports that recently emerged genes are expected to

display exaggerated genic features, rather than features intermediate between non genic and genic sequences (Wilson et al. 2017; Van Oss and Carvunis 2019). According to Wilson et al. (2017), novel genes are only born from sequences that happen to be pre-adapted, not to be harmful for the cell. Such non-harmful sequences are soluble sequences, with high intrinsic disorder propensity, permitting them to avoid deleterious non-specific aggregation in the aqueous cellular environment.

Analyzing mouse and yeast genes, they showed that younger genes are predicted to present higher propensity for intrinsic disorder compared with their older counterparts and random non-genic sequences as well (Wilson et al. 2017). This higher disorder propensity of young genes was initially attributed to their amino acid composition. Studying the clustering of hydrophobic amino acids on the mouse genes' sequences, Foy et al. (2019) observed that young genes show excess concentration of hydrophobic residues near one another while old genes present interspersion of their hydrophobic residues. They attributed this increased hydrophobic clustering of young proteins to a strategy of aggregation avoidance. These results support the important contribution of amino acid ordering together with the amino acid composition in young genes and could be seen as a preadaptation for de novo gene birth. The notion of preadaptation corresponds to specific characteristics (i.e., amino acid composition and ordering) of noncoding sequences which make them more favorable to give birth to de novo genes contrary to the large pool of all noncoding sequences. The model speculates that the gene birth is a sudden transition to functionality that occurs when an ORF acquires a selected effect (Van Oss and Carvunis 2019). Moreover, Bitard-Feildel et al. (2015) observed that orphan domains of *D. melanogaster* are likely to contain more large loops than ancient domains and that present a unique pattern of high intrinsic disorder and potential binding affinity. They speculated that this might correspond to sequences having a particular structural behavior, able to switch between ordered and disordered states.

Basile et al. (2017) supported that the opposite trends observed by these models of de novo genes emergence could be explained by the difference in GC content of the genome of the organisms. The claims of Wilson et al. (2017) were based on studies conducted on the mouse or fly proteins while the ones of Carvunis et al. (2012) were made on the yeast. Notably, the yeast genome is AT-rich (40% GC content) while the fly genome is GC-rich

(53% GC content) and the GC content has been proven to be correlated with codons coding for disorder promoting amino acids. More precisely, the codons encoding for Ala, Pro and Gly (disorder promoting amino acids) contain 80% GC while the ones encoding for Phe, Tyr, and Ile (order promoting amino acids) contain 20% or even less. They observed that low-GC orphans of *S. cerevisiae* were predicted as less disordered than the high-GC orphans of *D. melanogaster* while this relationship was weaker in the older proteins of both organisms. Notably, the structural properties of the youngest proteins (disorder propensity, content of secondary structure, fraction of transmembrane residues, fraction of low-complexity residues) resemble properties of random proteins with similar GC content. These results support that de novo created orphan proteins are very similar to random proteins respecting the general GC content of the organism while older proteins show lower dependency of their structural properties on GC. It must be mentioned that the GC content between young and older proteins remains the same though distinct for each organism. They speculate that selective pressure acts less on GC content and mostly on the structural features of proteins weakening their correlation through the evolutionary process. Interestingly, in line with the observation made by Carvunis et al. (2012), they also observed that the negatively charged residues (D and E) were more represented in older proteins (and notably more than expected) at any GC level suggesting a gradual increase of negative charges during evolution (Basile et al. 2017).

1.10 Open questions about the role of the noncoding genome

Even though controversial, these models support that the initial de novo peptide, once established, will further evolve towards a more canonical and well folded protein. As a result, they all give a central role to the fold potential of noncoding ORFs in the emergence of genetic novelty. Consequently, several questions can emerge about the foldability of the peptides potentially "encoded" by noncoding regions and the impact of the structural properties of the peptides that could emerge from noncoding regions in de novo gene birth and finally in genome evolution.

Indeed, it is well known that the noncoding genome corresponds to unevolved intergenic regions with random nucleotides' distribution. As a result, ORFs harboring in these

intergenic regions will code for peptides with different amino acid compositions from canonical proteins. Nevertheless, experimental data report evidence of widespread transcription and pervasive translation of intergenic ORFs, making us wonder about the structural properties and the fold potential of the peptides produced by the ensemble of the intergenic ORFs. More generally, the fate and the impact on the cell, of all the peptides coming from pervasive translation, remain unknown and deserve more attention. This opens the question of how the pervasive translation can be tolerated by the cell and shows the importance of investigating the structural properties of the potential peptides that could result from pervasive translation on the cell (i.e., being potential future de novo genes or not).

On the other hand, proteomes are characterized by a large structural diversity including disordered proteins, globular ones or transmembrane proteins which aggregate in solution while being able to fold in lipidic environments. Moreover, despite their complex evolutionary history, protein-coding genes have had a noncoding ancestral origin (Nielly-Thibault and Landry 2019). This permits us to ask the question whether the large structural diversity observed in the proteomes today could be already encoded by noncoding ORFs which gave rise to novel genes, or whether this important structural variability of the proteomes was acquired during evolution. If and how the amino acid compositions of the noncoding ORFs can account for the structural states observed in proteomes are crucial questions to understand the relationship, if any, between the noncoding genome and the protein structure universe.

Finally, some of the pervasively translated products would provide to the organism a selective advantage in specific environmental changes. The latter can be then further subjected to selective pressure and be established as novel genes. As a result, this motivates us to study the early stages preceding de novo genes birth and raises the question whether de novo genes, emerge from noncoding ORFs presenting specific sequence and structural properties compared with the overall pool of noncoding ORFs.
Answering these questions would permit us to estimate the extent to which the under-explored noncoding genome could produce novel protein bricks which can act as innovation tools capable at either giving rise to novel genes or being integrated into pre-

existing proteins and thereby, investigate the potential contribution of the noncoding genome in protein evolution and structural diversity.

1.11 Structural properties of random-unevolved sequences

In line with the above questions, there are studies which tried to investigate the structural properties of random short amino acid sequences (reminding the noncoding ORFs) (Davidson et al. 1995; Chiarabelli et al. 2006; Knopp et al. 2019). Recently, Tretyachenko et al. (2017) studied the fold potential of random-unevolved amino acid sequences and compared them with biological protein sequences. Their results showed that random sequences are predicted with comparable secondary structures occurrences with known biological proteins supporting that, structural motifs are not so difficult to be generated in a random way. In addition, they showed that random sequences, similarly to biological proteins, adopt a wide range of aggregation propensity containing low aggregation propensity like disordered proteins, high aggregation propensity like membrane proteins and intermediate aggregation propensity. Overexpression of some random sequences and analysis of their solubility showed that all the random sequences with low aggregation propensity were soluble whereas the same was true only for 30% of the sequences with high aggregation propensity. Notably, 75% of the remaining sequences, despite their intermediate aggregation propensity were also soluble. These results support that random amino acid sequences with low or even intermediate propensity to aggregate could, in many cases, be soluble and potentially tolerated by the cells. Interestingly, LaBean et al. (2011) showed experimentally that short (71 residues), unevolved random-sequence polypeptides, with amino acid composition close to natural globular proteins, are capable of forming secondary structure elements and consequently fold into a more dynamic molten globule conformation.

Langenberg et al. (2020) analyzed a large dataset of globular domains and observed that segments contributing to protein stability present high conservation of amyloidogenicity as well, suggesting that the evolutionary pressure towards the increase of protein stability will consequently increase their aggregation propensity. They support that amyloid and globular structures present intrinsic structural properties which are driven by similar physicochemical proclivities. These results are in line with the hypothesis for

an amyloid-driven origin of life which supports that amyloid peptides constitute the first self-replicating and catalytic molecules of life thus serving as an "ancestral fold" from which the globular protein universe could have emerged (Greenwald and Riek 2012; Langenberg et al. 2020).

It should also be mentioned that protein aggregation is related to protein concentration and as result overexpression experiments which, by definition, produce high quantities of protein copies could eventually lead proteins to aggregate. A study on the yeast proteome, revealed that highly abundant proteins presented multiple chaperone interactions which counterbalance the aggregation propensity of the proteins (Ibstedt et al. 2014). Notably, in their recent pre-print Tretyachenko et al. (2021) show that almost 40-50% of their random-sequences library is soluble and structured upon interaction with the DnaK chaperone supporting that the cellular context could provide further stabilization to the produced peptides. Overall, these studies, support that random noncoding peptides expressed in low levels, despite their non-optimized premature structures, could potentially be tolerated by the cell without generating aggregates. Furthermore, in the case that they get established as novel genes, they could continue to evolve towards more soluble well-structured proteins.

1.12 Some structural examples of de novo genes

Even though comparative genomics have permitted the detection and evolutionary study of numerous de novo genes, to date no experimental structure of any de novo protein has been reported (Bungard et al. 2017). However, some attempts of experimental structural characterization of de novo genes' proteins have been conducted.

BSC4 de novo gene

The yeast gene *BSC4* ("bypass of stop codon") constitutes a well-studied case of a protein coding gene that emerged de novo from an ancestral noncoding sequence and participates in DNA damage repair during stationary phase (Cai et al. 2008; Bungard et al. 2017). Bsc4 is a strictly orphan protein of *S. cerevisiae* species which is conserved among all its strains but no orthologous sequence exists in any other fungal species (Bungard et al. 2017). Indeed, Cai et al. (2008) used synteny and phylogeny methods and

showed that *BSC4* gene is homologous to an intergenic noncoding genomic region of other fungal species. Contrary to old proteins that fold into stable and specific globular structures the young Bsc4 tends to have "rudimentary" or "molten globule" characteristics lacking specific tertiary structure (Bungard et al. 2017). Notably, numerous biochemical analyses (i.e., native MS, far and near UV circular dichroism, thermal and chemical denaturation) give evidence that the Bsc4 protein is neither an Intrinsically Disordered Protein nor a well folded globular one. However, Bsc4 presents resistance to proteolysis suggesting strong intramolecular interactions and high level of order (Bungard et al. 2017). Bsc4 protein forms predominantly soluble oligomers rather than monomers even at modest concentration, suggesting that the oligomerization is important for establishing the protein's stability and solubility in the cellular environment. Until today, nobody has managed to determine the exact structure of Bsc4 through crystallization or nuclear magnetic resonance (NMR) and that reflects the unspecific character of this protein structure. Figure 1.7A represents in a very interesting metaphoric way the rudimentary fold of the Bsc4 protein, presenting the proteins as origami.

Gdrd de novo gene

The fruit fly gene gdrd (or goddard) emerged de novo from an intronic sequence at least 50 million years ago (Mya) at the root of *Drosophila* genus (Gubala et al. 2017). This means that gdrd constitutes a *Drosophila* taxonomically restricted gene and is found conserved in most fly species with only exception *D. willistoni*. gdrd is expressed specifically in the male reproductive tract and most likely participates in spermatid elongation during spermatogenesis being an essential gene for male fertility in *D. melanogaster* species. Like in Bungard et al. (2017), Lange et al. (2021) used biochemical methods (i.e. circular dichroism, NMR, thermal denaturation) and showed that the Gdrd protein of *D. melanogaster* although presents average biochemical properties, contains a stably folded core consisted by a principal alpha helix or coiled-coil conformation. At note, even though they could detect the general structural content of the protein, they did not manage to characterize its exact 3D fold experimentally. In addition, Gdrd appears to be soluble, does not form oligomers nor aggregates. Then, using comparative genomics and structural bioinformatics approaches they (i) reconstructed the ancestral sequences that preceded the emergence of gdrd gene and (ii) modeled the 3D structure of all the

Drosophila Gdrd proteins (together with their theoretical ancestral protein sequences) through an *ab initio* sequence fold approach (iii) and tested the structural stability of these models through molecular dynamics (MD) simulations. As observed in figure 1.7B all the Gdrd protein models present highly similar structures all containing a central long alpha helix while the rest of the protein seems quite variant. The MD simulations validated the highly stable character of the long central alpha helix as well as the fluctuant character of the rest of the protein. Together all these results support that the structure of the Grdr de novo protein has been largely conserved since its origin while the authors speculate that its essential functionality in *D. melanogaster* specifically might have arisen later, possibly through local structural modifications at the protein's termini.

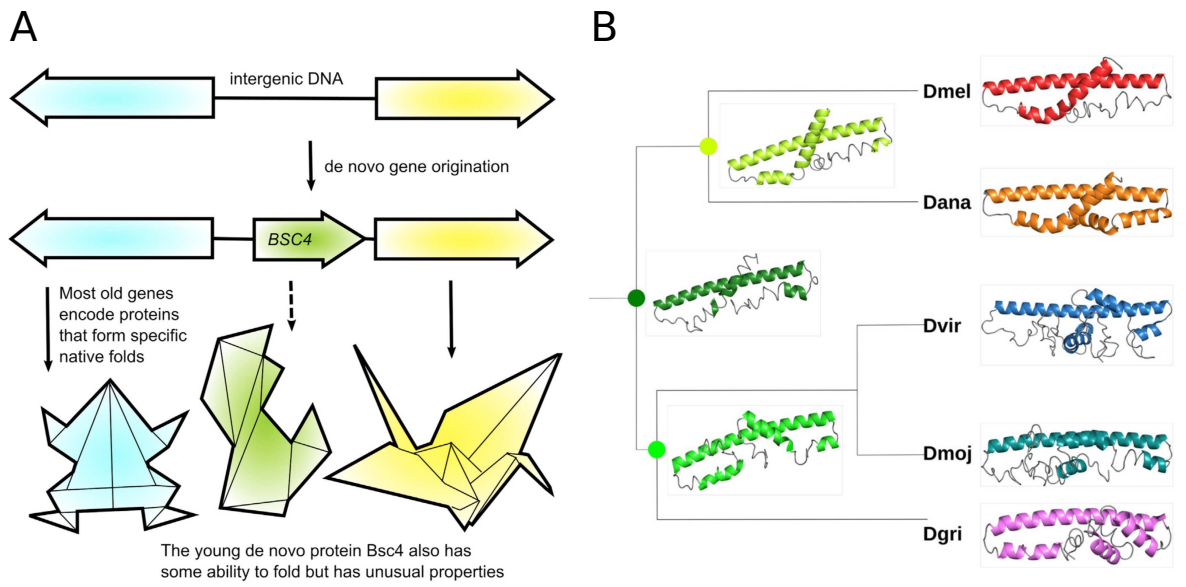


Figure 1.7. (A) Schematical and metaphorical representation as origami for the less specific folding of the de novo emerged Bsc4 protein (in green) compared with established proteins (in blue frog and yellow swan). The figure was extracted from the study of Bungard et al. (2017) with the title "*Foldability of a Natural De Novo Evolved Protein*". **(B)** Structural ab initio models of the de novo protein Gdrd of D. melanogaster (in red) and its orthologs in neighboring species (in orange, blue, turquoise and pink). Additionally, predictions for the most likely sequences for reconstructed ancestors of Dmel/Dana (bright green), Dvir/Dmoj/Dgri (green), and their most recent common ancestor (dark green) are shown (branch lengths are not meaningful). It is observed that all the protein models contain a central long alpha helix while the rest of the protein is more variant. The figure was extracted from the study Lange et al. (2021) with title "*Structural and functional characterization of a putative de novo gene in Drosophila*".

Overall, these two extremely interesting examples of de novo proteins converge towards the same observations. Young proteins that have emerged from ancestrally noncoding genomic regions present structural and biochemical characteristics comparable to the ones of conserved and well folded older proteins. However, these proteins do not appear to be uniquely folded into globular proteins but rather to a molten globule state described as a "rudimentary fold". All these observations enforce the hypothesis that de novo proteins capable of acquiring a fold, even partially, could potentially emerge from intergenic or intronic regions without prior adaptation (Lange et al. 2021). These novel proteins could constitute a source of genetic and structural innovation for the cell while after their birth and gain of expression their functionality and/or foldability could be adjusted through minor structural changes. Nevertheless, the highly dynamic structure of de novo proteins makes it difficult to characterize experimentally their capacity to fold and even more their exact 3D structure. As a result, structural computational biology seems to play an important role for the study of the fold potential and thereby the structure of novel proteins.

1.13 Predicting the fold potential of an amino acid sequence

The last few years, the prediction of the 3D structure for a given amino acid sequence has been a very intriguing challenge for the structural-computational biologists. Homology modeling methods are based to the assumption that similar amino acids sequences will fold into similar structures (Fiser and Šali 2003; Biasini et al. 2014; Schymkowitz et al. 2005) while protein threading methods try to superimpose an amino acid sequence on 3D scaffolds of known proteins (Söding et al. 2005; Peng and Xu 2011; Ghouzam et al. 2016). Then, the compatibility of the sequence to the structure is evaluated using physics-based scoring functions. Even though highly efficient, both these prediction methods are highly dependent on the sequence and structural information already existing in known databases making them less reliable for orphan sequences with no homologs (Kuhlman and Bradley 2019). On the other hand, the few *ab initio* methods existing are based mostly on structural "alphabets" demanding an exhaustive research of different structural "letters" combinations (Maupetit et al. 2009; Yang et al. 2020). Nevertheless, the computational demands of all three methods (but especially the *ab initio*) make them prohibitive for genomic scale analyses. Lately, the expansion of deep learning led to the

development of new and highly performant fold prediction methods (i.e., AlphaFold) which revisited the problem (Senior et al. 2020; Jumper et al. 2021). The new version of AlphaFold (AlphaFold²) was used to a large-scale study, providing full-length structure predictions for almost the entire human proteome (98.5% of human proteins) (Tunyasuvunakool et al. 2021). Strikingly, 58% of the total residues presented confident prediction with 36% presenting specifically very high confidence. Notably, all these methods aim at predicting a detailed structural model for a given amino acid sequence. However, the already presented examples of Bsc4 and Gdrd make it clear that young emerging proteins tend to fold into more dynamic structures with less defined structural content. Indeed, the prediction made on Bsc4 protein by AlphFold² reveals that the per-residue prediction confidence score is very low for the majority of the amino acids (Figure 1.8). Whether this low confidence score results from the highly dynamic character of its structure or simply from the fact that Bsc4 lacks homologous sequences which are important for the accuracy of AlphaFold² remains unknown.

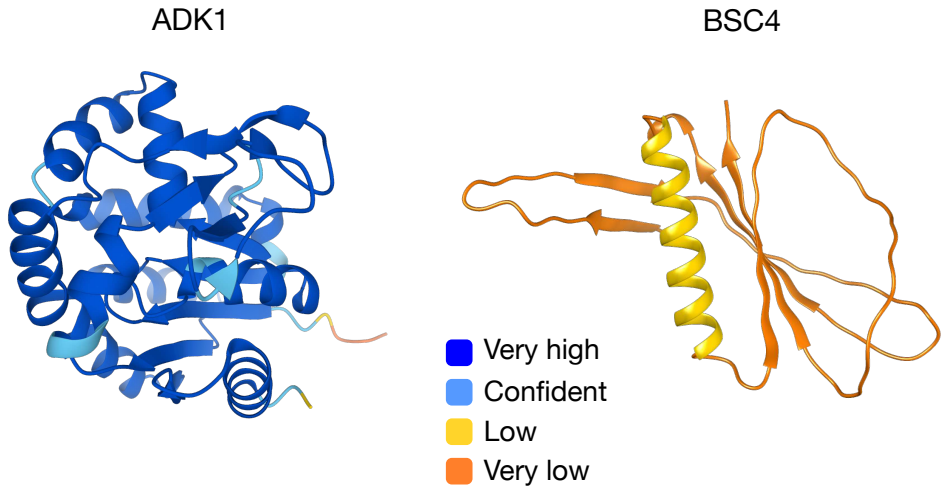


Figure 1.8. Prediction of the 3D structure of the Bsc4 and ADK1 proteins made by AlphaFold². The two models are colored based on the prediction confidence score of AlphaFold². Notably, the structure of adenylate kinase (ADK1) protein is predicted with high confidence (blue and light blue colored residues) while the Bsc4 protein is predicted with low confidence (yellow and orange residues).

The observations made on Bsc4 highlight two distinct but very important aspects of the problem. The first one corresponds to the prediction of the exact 3D fold of a protein while the second corresponds more basically at predicting the ability of a protein to fold in a given environment. These two questions demand different methods and provide different information. Indeed, predicting a detailed structural model for young orphan sequences whose capacity at acquiring a 3D fold is not well established would potentially lead to false predictions. Consequently, before the prediction of any detailed structural model, it becomes important to predict the ability of a given sequence to fold into a more or less stable fold (referred as foldability). The foldability of a protein informs us about the structural state of a protein rather than its exact 3D structure. Interestingly, the structural state of a protein can provide information on the potential behavior of a protein in the cell environment. Proteins can be completely disordered, folded in solution or can form aggregates in the cytosol, and will therefore behave differently according to their structural state.

In fact, differences between the fold state of proteins can be appreciated at the level of the amino acid sequence. Order-promoting residues, mostly corresponding to strong hydrophobic amino acids (V, I, L, M, Y, W, F), are known to participate in the formation of regular secondary structures and thus to the densely packed cores of globular domains and proteins. On the other hand, disordered regions are significantly depleted in order-promoting residues and enriched in disorder-promoting ones (A, R, G, Q, S, P, E, K). The Hydrophobic Clusters Analysis (HCA) constitutes an easily interpretable method which given the amino acid sequence of a protein, delineates clusters of strong hydrophobic amino acids (Bitard-Feildel et al. 2018). These hydrophobic clusters have been shown to be associated with regular secondary structures and are indicative of foldable domains. The hydrophobic clusters are connected by linkers corresponding to loops or disordered regions (Figure 1.9). The overall composition of hydrophobic clusters can provide information about the fold potential of the corresponding amino acid sequence. The fold potential of an amino acid sequence can be appreciated in a quantitative way with a foldability score where disordered sequences adopt mostly low values, sequences prone to fold in membranes and aggregate in solution adopt high values and sequences susceptible to fold adopt mostly intermediate values (Bitard-Feildel & Callebaut 2018). Analysis of real 3D folds (from the SCOP database) revealed that globular domains

present on average 33.3% of strong hydrophobic amino acids while membrane domains and peptides, around 41% with longer Hydrophobic Clusters. On the contrary, regions lacking Hydrophobic Clusters or containing small and scarcely distributed ones correspond mostly to highly disordered sequences or flexible linkers (Bitard-Feildel et al. 2018).

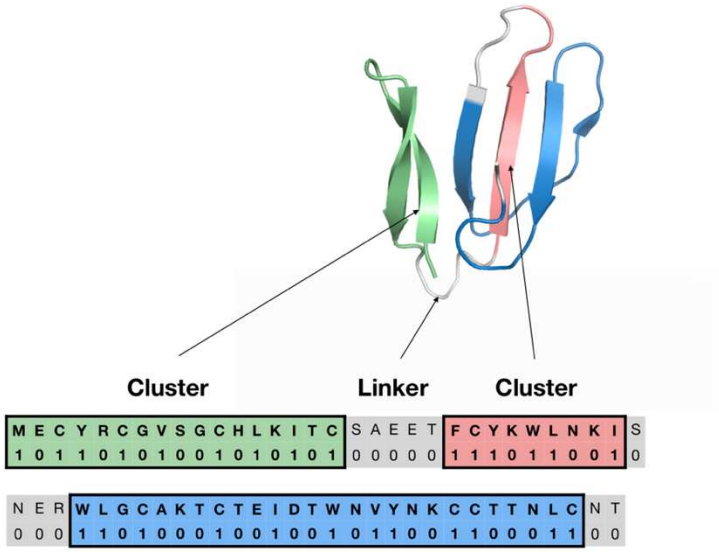


Figure 1.9. 3D mapping of HCA hydrophobic clusters and linkers. HCA hydrophobic clusters (colored) and linkers (in grey) delineated for the sequence of Bucandin (pdb code: 1f94). The HCA-based sequence, which consists in translating the protein sequence into a binary pattern, is given under the protein sequence. "1" corresponds to strong hydrophobic amino acids (V, I, L, F, M, Y, W) and "0" to the other amino acids. HCA clusters and linkers are mapped on the 3D structure of Bucandin with respect to the color code used for the sequence. Is interesting to observe that the hydrophobic clusters are not only the delimitations of regular secondary structures (helices and strands) but they can encapsulate more complex secondary structures' arrangements.

The advantage of HCA is that it is fast and needs the sole information of the amino acid sequence, without prior knowledge of any homologous sequence, thus offering a promising method to study the fold potential of orphan proteins (Bitard-Feildel et al. 2018). Moreover, HCA is a very fast, sequence-based foldability prediction method which makes it appropriate for large genome scale analyses. For example, among other studies,

HCA has been used for the detection of orphan domains in *D. melanogaster* proteome (Bitard-Feildel et al. 2015) or for the study of the foldability potential of the un-annotated part of the protein universe (referred as dark proteome) (Bitard-Feildel and Callebaut 2017).

1.14 Thesis' objectives

The general objective of my thesis was to study the potential role of the noncoding genome in the emergence of genetic novelty. In particular, I aimed at investigating how the noncoding genome participates in the emergence of de novo genes as well as in the evolution of proteins. In order to address this question, I adopted a structural point of view as it is well known that the functionality of the proteins is intimately related with their structure. Therefore, I characterized the fold potential diversity (i.e., propensity for disorder, folded state, or aggregation) of the amino acid sequences encoded by all intergenic ORFs (IGORFs) of *S. cerevisiae* in order to (i) estimate the potential of the noncoding genome to produce novel protein bricks, that can either give birth to novel genes or be integrated into pre-existing proteins, thus participating in protein structure evolution and diversity, and (ii) explore whether the large structural diversity observed in proteomes is already present in noncoding sequences, and thereby investigate the relationship, if any, between the fold potential of the amino acid sequences encoded by IGORFs and the structural diversity of proteins.

The first part of my thesis constituted at developing a bioinformatic method for the detection of all the IGORFs of *S. cerevisiae* and the estimation of the fold potential and other sequence and structural properties of the potential peptides encoded by them. For that purpose, I participated in the development of a bioinformatic tool called ORFtrack which aims at "tracking" all the ORFs of a given genome and annotate their overlapping (or not, in the case of IGORFs) with genomic annotated features. Then I developed ORFold which aims at estimating the fold potential as well as the disorder and aggregation propensity of a given amino acid sequence and I applied it on the peptides encoded by IGORFs. ORFold makes use of three academically free bioinformatic tools (pyHCA (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2018a; Lamiable et al. 2019), IuPRED2 (Mészáros et al. 2018; Dosztányi 2018; Erdős and Dosztányi 2020) and Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a), respectively) and gives an indication of the overall foldability (together with complementary information of the disorder and aggregation propensity) for every amino acid sequence encoded by IGORFs. These two bioinformatic tools are grouped together into one package called ORFmine which is freely accessible via GitHub (GitHub 2021). The package ORFmine together with a detailed step-by-step protocol of IGORFs

extraction and their peptides' fold potential estimation were presented in the book Methods Molecular Biology, in a special issue on "Computational Peptide Science" in the chapter entitled "*Exploring the peptide potential of genomes*".

I then characterized the early stages preceding de novo gene emergence with two complementary approaches (i) the systematic reconstruction of the ancestral noncoding sequences of 70 *S. cerevisiae* de novo genes in order to identify the sequence and structural features of IGORFs that indeed gave birth to known de novo genes and (ii) the identification of IGORFs with a strong translation signal through ribosome profiling experiments, in order to investigate the sequence and structural properties of candidate IGORFs that could give birth to future novel genes. At this part of my thesis, I developed a pipeline which permits to correctly map Ribosome Profiling data on noncoding sequences in order to detect IGORFs with translation signal. This pipeline called ORFribo will soon be part of the ORFmine package proposing a complete protocol for (i) IGORFs detection and extraction, (ii) prediction of the overall foldability potential of their amino acid sequences, and (iii) identification of interesting IGORF candidates presenting translation signatures. All the results about the fold potential and sequence and structural properties of peptides encoded by IGORFs are presented in a research article entitled "Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution" and which has been published at the Genome Research peer-reviewed journal (Papadopoulos et al. 2021).

In the next part of my thesis, using phylostratigraphy approaches I divided the *S. cerevisiae* proteins according to their relative phylogenetic age in order to investigate how fast are fixed and how evolve the sequence and structural properties of the yeast proteins along the evolutionary time.

In the final part of my thesis, I developed a supervised machine learning model which aims at predicting the folding state (i.e., disordered, stable in solution, stable upon interaction with a partner or transmembrane) of the potential peptides encoded by IGORFs and therefore predict their potential "behavior" in the cellular environment. The objective of this part was to explore more finely the structural properties of the peptides

encoded by IGORFs in order to better understand the emergence of de novo genes and further investigate how the pervasive expression of IGORFs could be tolerated by the cell.

Despite its compact genome, we selected *S. cerevisiae* for our study because corresponds to a eukaryotic model organism whose genome has been completely sequenced and well annotated thus, permitting to identify with high confidence the intergenic regions and consequently the IGORFs. In addition, different teams in our institute, with which we collaborate, work on *S. cerevisiae* and they could provide us with experimental data (i.e., Ribosome Profiling) permitting us to detect interesting IGORF candidates with experimental proof of expression. Furthermore, the genomes of species closely related to *S. cerevisiae* are also available, permitting us to detect the noncoding genomic regions of known de novo genes of *S. cerevisiae* and thus reconstruct their ancestral sequences.

2 Methodology

In this section are presented the principal methods that were developed during my thesis, in order to:

(i) extract all the ORFs of a genome and annotate them according to specific genomic features (i.e., ORF overlapping a lncRNA, a protein coding gene, intergenic ORF etc.),

(ii) investigate the fold potential and other structural properties of each ORF being coding or not,

(iii) probe the translation activity of all ORFs.

Therefore, we created a freely distributed package named ORFmine which consists of two bioinformatic tools. The first one, named ORFtrack, was developed by my colleague Nicolas Chevrollier and aims at "tracking" all the ORFs of a genome (with a STOP-to-STOP ORF definition) and annotate them based on their overlapping with known annotated genomic features. At note, ORFtrack adopts an ORF-centered point of view of the genome and ORFs do not correspond to real biological objects but mostly reflect the potential peptides that could be expressed from a genome .The second program, called ORFold and developed by me, aims at predicting the foldability potential together with the disorder and aggregation propensity of any amino acid sequence by combining the results of three independent bioinformatic tools: pyHCA (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2018a; Lamiable et al. 2019), IuPRED2 (Mészáros et al. 2018; Dosztányi 2018; Erdős and Dosztányi 2020) and Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a), respectively. The two principal tools of ORFmine (ORFtrack and ORFold) are completely independent permitting them to be integrated in different protocols. However, combined together they offer to the user a complete analysis of the fold potential of both coding and noncoding ORFs of any genome. A step-by-step protocol for the use of ORFmine with numerous examples on different organisms was presented in the book Methods Molecular Biology, in a special issue on "Computational Peptide Science" in the chapter entitled "*Exploring the peptide potential of genomes*".

In addition, with the help of a master student who I supervised, Camille Rabier, we developed a pipeline (named ORFribo) which aims at mapping Ribo Seq data on the intergenic regions and detecting the frame under translation among the three possible ones. ORFribo, is not yet integrated in the ORFmine package as it still needs some adjustments, but its protocol will be presented in this chapter.

2.1 Exploring the peptide potential of genomes

Chris Papadopoulos[1], Nicolas Chevrollier[2], Anne Lopes[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France
[2] Independent investigator

Abstract

Recent studies attribute a central role to the noncoding genome in the emergence of novel genes. The widespread transcription of noncoding regions and the pervasive translation of the resulting RNAs offer to the organisms a vast reservoir of novel peptides. Although the majority of these peptides are anticipated as deleterious or neutral and thereby, expected to be degraded right away or short-lived in evolutionary history, some of them can confer an advantage to the organism. The latter can be further subjected to natural selection and be established as novel genes. In any case, characterizing the structural properties of these pervasively translated peptides is crucial to understand (i) their impact on the cell and (ii) how some of these peptides derived from presumed noncoding regions can give rise to structured and functional de novo proteins. Therefore, we present a protocol that aims to explore the potential of a genome to product novel peptides. It consists in annotating all the open reading frames (ORFs) of a genome (i.e. coding and noncoding ones), and characterizing the fold potential and other structural properties of their corresponding potential peptides. Here, we apply our protocol to a small genome and then show how to apply it to very large genomes. Finally, we present a case study which aims to probe the fold potential of a set of 721 translated ORFs in mouse lncRNAs identified with ribosome profiling experiments. Interestingly, we show that the distribution of their fold potential is different from the one of the nontranslated lncRNAs and more generally from the other noncoding ORFs of the mouse.

Running Head: Mining noncoding genomes

Key words: noncoding DNA, fold potential, *de novo* genes, small ORF-encoded peptides, ORFtrack, ORFold

1 Introduction

Many studies attribute a central role to the noncoding genome in novel gene birth and more generally in the emergence of genetic novelty. As a matter of fact, thousands of small Open Reading Frames (ORFs) have been identified in noncoding regions of various genomes. Interestingly, the wide use of transcriptomics revealed a high pervasive transcription of noncoding regions, and an important fraction of the resulting RNAs have been shown to be translated by ribosome profiling experiments (Ingolia et al. 2011; Ruiz-Orera et al. 2018; Li and Liu 2019; Chen et al. 2020). In addition, mass spectrometry experiments conducted on mammals, bacteria, or plants (Samayoa et al. 2011; Hobbs et al. 2011; Slavoff et al. 2013; Prabakaran et al. 2014; Eguen et al. 2015; Deng et al. 2018; Wang et al. 2019), confirm the existence of these translation products in the cell with the identification of hundreds of peptides derived from noncoding regions. The fact that these noncanonical products display short sizes, are present in low abundance, and use alternative start codons renders difficult their identification and suggests that their number is largely underestimated. Interestingly, their sequences are more conserved than those of noncoding sequences suggesting that they are subjected to purifying selection (Slavoff et al. 2013; Prabakaran et al. 2014) and they could be functional. It has been proposed that these noncanonical translation products are consequently exposed to natural selection and thereby, provide the organism with the raw material for the emergence of genetic novelty. However, how noncoding sequences can give rise to novel genes remains unclear. Particularly, noncoding sequences are not expected to fold to a stable and specific structure and have not been subjected to purifying selection in order not to be deleterious for the cell. One can ask how these pervasively translated products can (i) be tolerated by the cell and (ii) give rise to functional products, since most proteins achieve their function through a well-defined 3D structure. Indeed, noncoding sequences display different sequence features from coding ones, being shorter and characterized by different nucleotide compositions (Carvunis et al. 2012; Slavoff et al. 2013). They are rather expected to encode disordered, misfolded, or aggregation-prone peptides and we can hypothesize that they would be rapidly degraded or short-lived in evolutionary history. Nevertheless, it has been demonstrated that proteins from random libraries could fold *in silico* or *in vitro*, some of which being even beneficial in *Escherichia coli* (Keefe and Szostak 2001; Schaefer et al. 2010; Tretyachenko et al. 2017; Neme et al.

2017). All these results place the foldability of noncoding ORFs at the center of novel gene birth and strengthen the need for the characterization of the fold potential (including the propensities for disorder, folded state, and aggregation) not only of the experimentally observed de novo peptides but also of all the amino acid sequences "encoded" by presumed noncoding ORFs which could give rise to novel peptides upon pervasive translation.

Therefore, we present a protocol that enables in an automated way (i) the extraction and annotation of all possible ORFs of a genome, and (ii) the prediction of their fold potential along with their propensities for disorder and aggregation. It relies on the ORFmine package (unpublished but available at https://github.com/i2bc/ORFmine) which aims to annotate a genome's ORFs and probe their fold potential and structural properties. ORFmine consists of two independent programs ORFtrack and ORFold. ORFtrack works in a stand-alone fashion and is very flexible, enabling different levels of annotation depending on the user request. ORFold relies on three gold-standard programs, HCA (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2018a; Lamiable et al. 2019), Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a), and IUPred2A (Mészáros et al. 2018; Dosztányi 2018; Erdős and Dosztányi 2020) which predict respectively the fold potential, the aggregation, and the disorder propensities of an amino acid sequence. Here, we consider as foldable, the amino acid sequences which are able to fold to a stable 3D structure or to a molten globule state in which the specific tertiary structure is lost, whereas the secondary structures are intact. Our protocol can be applied to any completely sequenced genome and takes a few hours on a personal computer for a small genome (bacteria, archaea, or fungi), although we recommend launching the pipeline on a cluster for larger genomes (e.g. plant or mammal genomes). Here we present a detailed application of our protocol on the small genome of *Escherichia coli*. Then we show how to apply our protocol to very large genomes (*Mus musculus*). In the last part, we present a case study based on a ribosome profiling experiment performed on the mouse. In this example, we probe the fold potential of 721 ORFs present in lncRNAs which are translated, not conserved across species and which show weak or no signature of selective pressure (i.e. presumed as noncoding). We then show how ORFold can be used to compare the fold potential of a subset of ORFs of interest (e.g. translated ORFs present in lncRNAs) with those of the coding and noncoding ORFs of the genome they belong to. The latter protocol can be extended to any set of sequences

of interest including for example, peptides identified in mass spectrometry experiments carried out in different conditions, de novo peptides associated with specific diseases or even designed sequences.

2 Materials

2.1 ORFmine

ORFmine is a package that we developed in order to explore the peptide potential of a noncoding genome with the extraction and the annotation of all the possible ORFs present in noncoding regions. The ORFmine package is not published yet but available at: https://github.com/i2bc/ORFmine and consists of two independent programs, ORFtrack and ORFold that can be combined together or used independently (Figure 2.1). Used together, ORFtrack and ORFold provide a global picture of the fold potential and the structural properties of all the potential peptides of a genome. Otherwise, ORFtrack can simply be used to extract and annotate the ORFs of a genome, while ORFold can estimate the fold potential of any set of sequences without using genomic information.

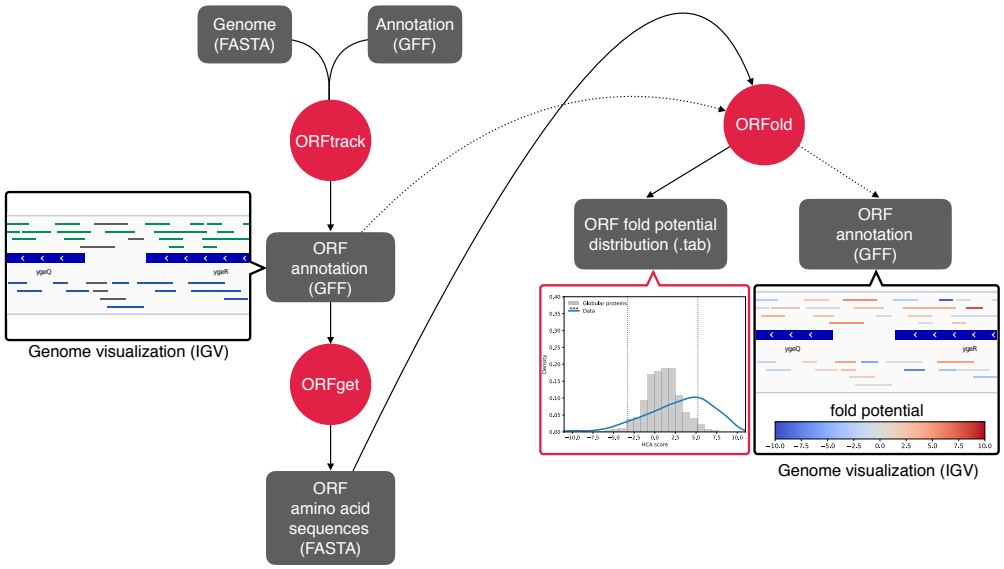


Figure 2.1: Pipeline of ORFmine. The inputs and outputs are represented with grey rectangles while the main scripts are shown with red circles. The mandatory inputs necessary to the ORF annotation and the estimation of their structural properties (e.g. fold potential and disorder and aggregation propensities), as

well as their corresponding outputs are connected to their related scripts with black arrows. The classical pipeline of ORFmine provides the user with a plot representing the distribution of the fold potential of the input ORFs (red box). Optionally, a genome annotation file (GFF format) can be given to ORFold (dashed arrows). In this case, ORFold produces new GFF files (one per studied structural property) where all input ORFs are associated with the score of the corresponding property. The GFF produced by ORFtrack and ORFold can be subsequently uploaded on a genome viewer (black boxes) where ORFs will be colored according to their annotation (black box on the left) or their structural properties (black box on the right).

2.1.1 ORFtrack

ORFtrack aims at extracting and annotating all the possible ORFs of a genome according to a set of defined genomic features. It takes as inputs a FASTA file containing all the chromosome or contig sequences and its corresponding annotation GFF file (for more details, see the GFF3 file format description at https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md). ORFtrack searches, in the six possible frames, for all possible ORFs of at least 60 nucleotides bounded by STOP codons (i.e. it does not search for start codons). In order to annotate each resulting ORF (e.g., intergenic ORF, noncoding ORF that overlaps a coding sequence, coding ORF etc.), their localization is subsequently compared to those of all genomic features annotated in the GFF file (e.g. CDS, tRNA, rRNA, or any other feature defined by the user in the third column of the GFF file) (Figures 2.2-2.3). There are four main categories of ORFs: (1) Coding ORFs (c_CDS) which correspond to ORFs that include a coding sequence (CDS) (i.e., in the same frame of a CDS). They are generally larger than the CDS since they are defined STOP-to-STOP. (2) Noncoding intergenic ORFs (nc_intergenic) which do not overlap any genomic feature. (3) Noncoding ORFs which overlap a genomic feature on the same strand (nc_ovp_same-x with x standing for the corresponding genomic feature), and (4) Noncoding ORFs which overlap a genomic feature on the opposite strand (nc_ovp_opp-x with x standing for the corresponding genomic feature) (Figures 2.2-2.3).

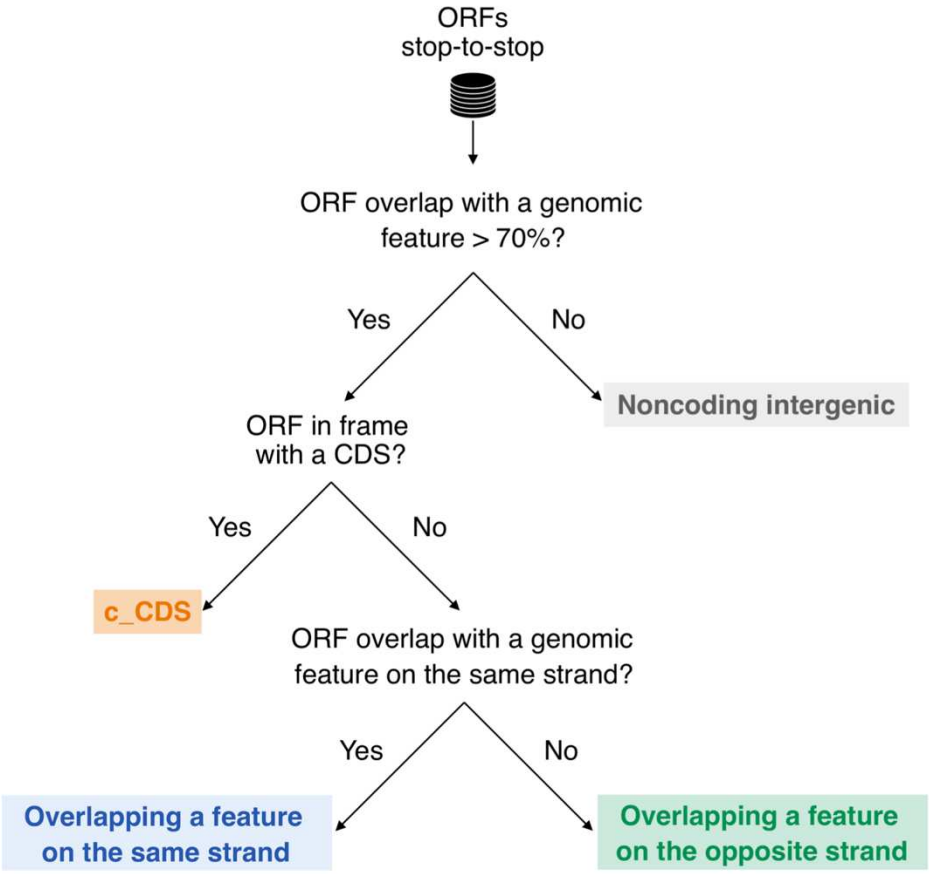


Figure 2.2: Decision tree of ORFtrack. ORFs are annotated according to four main categories: c_CDS for coding ORFs (orange box), noncoding intergenic ORFs (grey box), noncoding ORFs that overlap a genomic feature on the same strand (blue box) or on the opposite strand (green box).

The user has to keep in mind that ORFtrack provides an ORF-centered point of view of the input genome and that ORFs do not correspond to real biological objects but rather to the potential peptides that could be produced upon pervasive translation with no information on the localization of their first translated codon. For example, a noncoding ORF overlapping a tRNA does not correspond to a tRNA which by definition has neither phase nor a corresponding amino acid sequence, but to the corresponding peptide which could be produced upon the pervasive translation of the tRNA gene with no knowledge of the first translated codon.

If a noncoding ORF overlaps more than one genomic feature, ORFtrack applies the following priority rules:

- the noncoding ORF overlaps a CDS and any other genomic feature: it is annotated as a noncoding ORF overlapping a CDS (same or opposite strand) (e.g. nc_ovp_[same/opp]-CDS)
- the noncoding ORF overlaps a genomic feature on the same strand and any other genomic feature on the other strand (except CDS): it is annotated as a noncoding ORF overlapping the feature on the same strand (e.g. nc_ovp_same-x)
- the noncoding ORF overlaps two or more genomic features located on the same strand that can correspond to the same or the opposite strand of the noncoding ORF: it is annotated as overlapping the genomic feature that has the larger overlap with it (e.g. nc_ovp_[same/opp]- x)

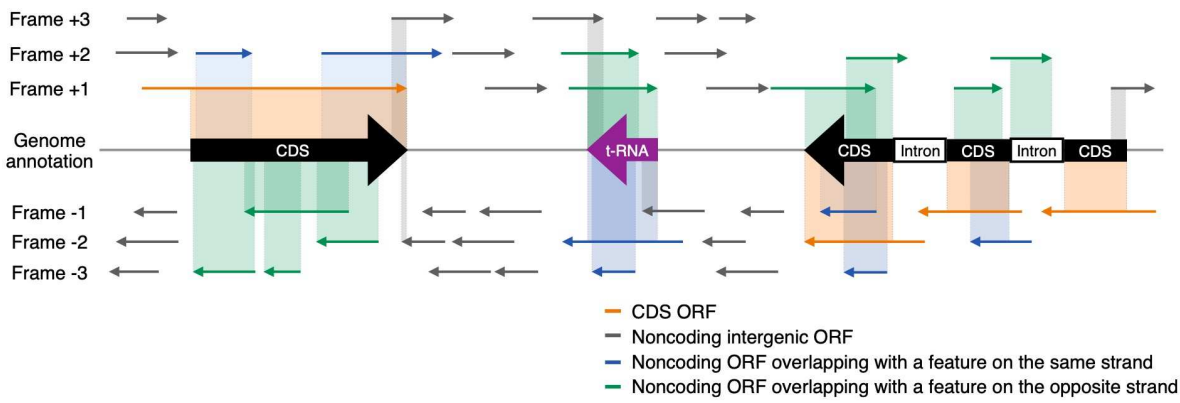


Figure 2.3: Schematic representation of the six frames of a DNA section. The genomic features annotated in the original GFF file are represented in the middle line. The ORFs of the six frames are colored with respect to their ORFtrack annotation. The overlap between an ORF and a genomic feature is illustrated with a rectangle colored according to the ORF annotation.

The program provides the user with a new GFF file containing all the identified ORFs annotated according to the four categories defined previously. ORFget (a tool provided with ORFtrack) generates a FASTA file containing the amino acid sequences of all identified ORFs or a subset of ORFs selected with respect to their annotation category (e.g. c_CDS, nc_intergenic, nc_ovp_same, nc_ovp_opp) or to their complete annotation for a finer selection (e.g. nc_ovp_same-lncRNAs and nc_ovp_opp-lncRNAs if, for example, the user seeks to investigate whether ORFs overlapping lncRNAs display specific properties

compared to other noncoding ORFs - see Subheading 3.3 for an example). Finally, ORFget allows the user to extract in a FASTA file, the amino acid sequences of all annotated proteins, and to reconstruct all isoforms of multi-exonic genes if they are annotated in the input GFF file.

2.1.2 ORFold

ORFold aims at estimating the fold potential of a set of amino acid sequences using the HCA method (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2018a; Lamiable et al. 2019). In addition, it can predict their disorder or aggregation propensities with IUPred (Mészáros et al. 2018; Dosztányi 2018; Erdős and Dosztányi 2020), and Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a), respectively. Although HCA is very fast and can handle all ORFs of a small genome in a few minutes, the calculation of the disorder and aggregation propensities slows down ORFold (around 3 hours on a single CPU (2GHz processor, 16GB RAM) for all the ORFs of *Escherichia coli*). Consequently, the user can turn off the calculation of the disorder and aggregation propensities. ORFold takes as input a FASTA file containing the amino acid sequences to treat. The output of ORFold is a table containing the fold potential and/or the disorder and aggregation propensities of each input sequence. Optionally, the user can provide ORFold with the genome annotation GFF file of the input genome. In this case, the fold potential and/or the disorder and aggregation propensities of each ORF will be added in the GFF file. The latter can be uploaded subsequently on a genome viewer such as IGV (Robinson et al. 2011), enabling the visual inspection and manual analysis of the distribution of the fold potential and the other structural properties along the genome. The program can handle several FASTA files at the same time and will generate as many outputs as given FASTA files. Finally, ORFold can also provide the user with plots representing the distribution of the fold potential of the input sequences along with those of a dataset of globular proteins used as reference taken from Mészáros et al. (2018).

HCA

ORFold estimates the fold potential with the HCA (Hydrophobic Cluster Analysis) approach (Bitard-Feildel and Callebaut 2017, 2018b). HCA toolkit is available at https://github.com/T-B-F/pyHCA. It splits an amino acid sequence into hydrophobic clusters and linkers. The formers gather strong hydrophobic residues (V, I, L, F, M, Y, W)

and cysteines while the latter correspond to stretches of residues which are composed of at least four non-hydrophobic residues or a proline. Hydrophobic clusters usually indicate one or several regular secondary structures connected by short loops, which constitute signatures of globular domains. Linkers correspond to loops or disordered regions. The fold potential of a sequence is determined by its composition in hydrophobic clusters and linkers and is reflected with the HCA score. The latter ranges from -10 to +10 with low HCA scores indicating sequences that are enriched in linkers and expected to be disordered. High HCA scores correspond to sequences with a high density in hydrophobic clusters, and are likely to form aggregates in solution, though some of them may be able to fold in lipidic environments. Sequences that are able to fold in solution are usually characterized by intermediate HCA scores as shown with the HCA scores of the reference dataset of globular proteins in Figure 2.5.

Tango

ORFold calculates the aggregation propensity of a sequence with Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a) which is available here http://tango.crg.es upon request from the developers. Following the criteria proposed by Linding et al. (2004), a sequence segment is considered as aggregation prone if it is composed of at least five consecutive residues predicted as populating a b-aggregated conformation with a percentage occupancy greater than 5%. The aggregation propensity of a sequence is then calculated as the fraction of residues predicted in an aggregation prone segment.

IUPred

ORFold calculates the disorder propensity with IUPred (Mészáros et al. 2009, 2018; Dosztányi 2018; Erdős and Dosztányi 2020). We use the version 2A of IUPred (Mészáros et al. 2018; Erdős and Dosztányi 2020) which is available here https://iupred2a.elte.hu upon request from the developers. Consistently with the criteria used for the definition of an aggregation prone region, we considered as disordered, a region composed of at least five consecutive residues displaying a disorder probability higher than 0.5. According to the aggregation propensity calculation, the disorder propensity of a sequence is calculated as the fraction of residues predicted in a disordered prone segment.

3 Methods

3.1 Classical use: probing the fold potential of a complete genome

Here we seek to probe the fold potential and the aggregation and disorder propensities of all noncoding ORFs of *Escherichia coli* str. K-12 substr. MG1655 (*E. coli*) regardless they overlap a genomic feature. As a reference, we will also characterize these properties for all CDS of *E. coli*.

3.1.1 FASTA and GFF files used in this example

- E_coli.fna (available at https://github.com/i2bc/ORFmine in the "examples" directory)
- E_coli.gff (available at https://github.com/i2bc/ORFmine in the "examples" directory)

3.1.2 Annotation of the ORFs of *E. coli* with ORFtrack

The following instruction of ORFtrack displays all the genomic features annotated in the genome of *E. coli*:

```
> orftrack -fna E_coli.fna -gff E_coli.gff --show-types
```

Up to 12 different genomic features are annotated in the *E. coli* genome including CDS, tRNA, rRNA... (see Note 1). We then annotate all the possible ORFs of *E. coli* with the following instruction:

```
> orftrack -fna E_coli.fna -gff E_coli.gff
```

The execution time on a single CPU (2GHz processor, 16GB RAM) is 38 seconds. ORFtrack generates a new GFF file (mapping_orf_E_coli.gff) that contains 135097 annotated ORFs among which 130637 are annotated as noncoding. Table 1 shows the distributions of the output ORFs across the different annotation categories with various levels of annotations. This information is available in the summary file produced by ORFtrack (summary.log). Notice that it is also possible to scan all the annotated ORFs by loading the new GFF on a genome viewer.

Table 2.1: Counts of *E. coli* ORFs for each annotation category

Total ORFs				
135097				
Coding (c_CDS)	Noncoding (nc_*)			
4460	130637			
	Noncoding intergenic (nc_intergenic)	Noncoding overlapping with a genomic feature… (nc_ovp_*)		
	18318	112319		
		On the same strand (nc_ovp_same-x)		On the opposite strand (nc_ovp_opp-x)
		47880		64439
		with x standing for:		
		45053	CDS	62354
		1136	repeat region	545
		626	sequence feature	566
		607	r-RNA	528
		140	nc-RNA	130
		119	t-RNA	114
		119	pseudogene	109
		77	mobile genomic element	87
		3	origin of replication	4
		0	recombination feature	2

3.1.3 Extraction and writing of the noncoding ORFs and the CDS of *E. coli*

Extraction of noncoding ORFs

In this example, we consider all the 130637 noncoding ORFs and do not differentiate noncoding intergenic ORFs from those that overlap a genomic feature. Therefore, we extract and write the amino acid sequences of all noncoding ORFs (i.e. nc_intergenic, nc_ovp_same and nc_ovp_opp) with ORFget with the following command line (see Note 2):

```
> orfget -fna E_coli.fna -gff mapping_orf_E_coli.gff -
features_include nc -o E_coli_noncoding
```

ORFget generates a FASTA file with the resulting 130637 amino acid sequences.

Extraction of CDS

Finally, in order to compare the structural properties of CDS with those of the potential peptides "encoded" in noncoding regions, we extract and rebuild the amino acid sequences of each CDS of *E. coli* according to the original annotation GFF file.

```
> orfget -fna E_coli.fna -gff E_coli.gff -features_include CDS -o
E_coli_CDS
```

We obtain a FASTA file of 4316 protein sequences.

3.1.4 Characterization of the fold potential, and the disorder and aggregation propensities of the ORFs and CDS of *E. coli* with ORFold

We aim at characterizing the fold potential, and the disorder and aggregation propensities of the noncoding ORFs (intergenic and overlapping ORFs) and CDS of *E. coli*. ORFold can handle the two datasets at the same time with the following instruction:

```
> orfold -fna E_coli_noncoding.pfasta E_coli_CDS.pfasta -gff
mapping_orf_E_coli.gff E_coli.gff - options HIT
```

The execution time on a single CPU is around 3 hours. ORFold generates two tables (one per dataset) containing for each sequence, its fold potential, as well as its disorder and aggregation propensities calculated by HCA, IUPred and Tango, respectively. In addition, ORFold writes the output values in a new GFF file that can be uploaded on a genome viewer. The original GFF can be uploaded as well, providing a reference with the exact localization of the genomic features annotated in the original GFF. We recall that ORFtrack identifies and annotates all the possible ORFs of a genome which do not correspond to real objects but rather to the potential peptides that could be produced if

their corresponding DNA region is transcribed, and the resulting RNA subsequently translated.

Figure 2.4 shows the two DNA strands of a genomic section of *E. coli* represented by the genome viewer IGV (Robinson et al. 2011) after uploading the original GFF (blue genes in the middle) and the new GFF returned by ORFtrack (small ORFs in the panels 2 and 4). Although the genome of *E. coli* is very compact with a few intergenic regions, there is a high density of noncoding ORFs that overlap with the coding genes of *E. coli*, and that represent a high potential of novel peptides in case of ribosomal frameshifting. Interestingly, the distribution of the fold potential along the genome is not homogeneous. We observe an island of noncoding ORFs with high HCA values (ORFs in light and dark red in the middle of the figure). These ORFs potentially encode peptides enriched in hydrophobic residues that are likely to be foldable (light red ORFs) or expected to form aggregates in solution (dark red ORFs). The GFF returned by ORFold containing the Tango or IUPred values can provide the user with complementary information (data not shown). The genomic regions around the island of high HCA values ORFs, are enriched in ORFs with intermediate HCA values typical of foldable sequences (ORFs in light red and light blue). Overall, it is interesting to note that the fold potential seems to be quite conserved among the three frames of a strand, though it can vary along the strand. This recall the observation made by Bartonek et al. (2020), who showed that the hydrophobicity profiles of protein sequences are preserved in +1, -1 frames through the structure of the genomic code. Finally, the visual inspection of the distribution of the fold potential of noncoding ORFs suggests that there is a vast amount of ORFs that potentially encode foldable peptides (light blue and light red boxes corresponding to intermediate HCA values). Whether these peptides would fold to a specific 3D structure or to a molten globule is a crucial and very difficult question that deserves further investigation.

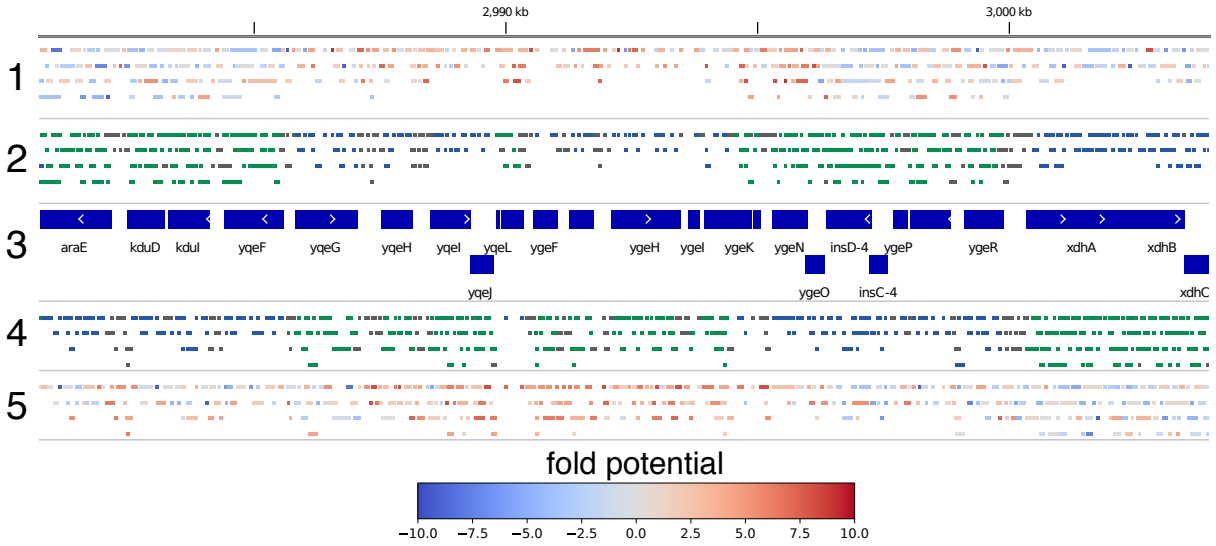


Figure 2.4: Screenshot of a genomic section of *E. coli* represented by IGV. Genomic features present in the original GFF file (CDS in this example) are represented with blue boxes in the middle of the figure (panel 3). Panels 2 and 4 represent the noncoding ORFs identified by ORFtrack in the positive and negative strands, respectively. They are colored according to their annotation category (grey, blue and green for nc_intergenic, nc_ovp_same and nc_ovp_opp respectively). Panels 1 and 5 represent the same ORFs colored with respect to their HCA scores. ORFs with low HCA scores are colored in blue, whereas ORFs with high HCA scores are colored in red. For more clarity c_CDS which correspond to ORFs including a CDS in the same frame are not shown since the corresponding CDS are already represented with the blue boxes in the middle panel.

Finally, we plot the distributions of the fold potential of the two datasets with ORFplot. Notice that ORFplot can deal with several inputs and will plot as many distributions as given tables.

```
> orfplot -tab E_coli_CDS.tab E_coli_nocoding.tab -names “E. coli CDS” “E. coli noncoding ORFs”
```

Figure 2.5 shows the fold potential distributions of the noncoding ORFs and the CDS of *E. coli* as plotted by ORFplot. Furthermore, as a reference, ORFplot plots the distribution of the HCA scores of a set of globular protein sequences taken from Mészáros et al. (2018). The fold potential distribution of the CDS is clearly different from the one of the noncoding sequences (KS test, P = 9.9e-18). The CDS are enriched in intermediate HCA

values typical of foldable proteins as shown by the HCA scores of the globular proteins. Conversely, noncoding ORFs display a wide range of HCA values reflecting foldable, disordered or aggregation prone potential peptides. Nevertheless, it is interesting to note that the majority of them (~64%) exhibit similar HCA scores to globular proteins, revealing an important potential of foldable peptides in line with the observation made in Figure 2.4.

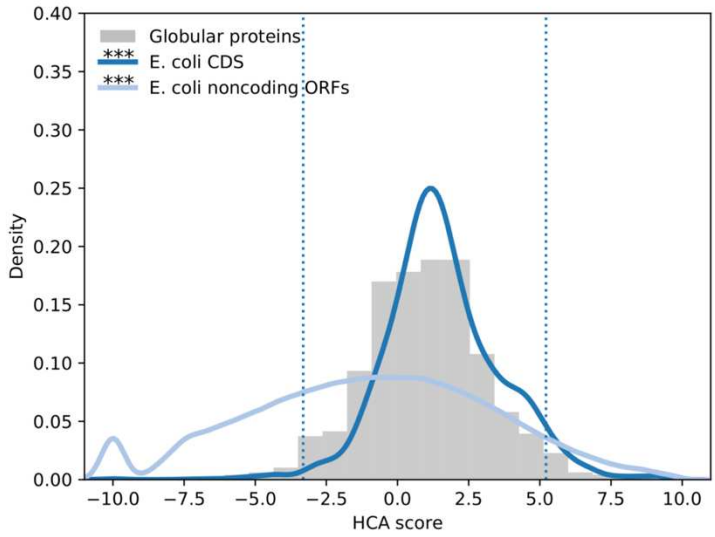


Figure 2.5: Distribution of the HCA scores calculated for the CDS and the noncoding ORFs of *E. coli* (dark blue and light blue curves respectively). The HCA score distribution of the set of globular proteins is represented by the grey histogram. Dotted black lines delineate the boundaries of the low, intermediate and high HCA score bins so that 95% of the globular proteins fall into the intermediate HCA score bin. Each distribution is compared with the one of the globular proteins set with a Kolmogorov Smirnov test. Asterisks on the plot denote level of significance: *** < 1×10^{-3}.

3.2 Application to large genomes and comparison with other species

The execution time and the size of the outputs increase with the size of the input genome. This can become dramatical for very large genomes such as those of mammals or plants. Even if the execution time for ORFtrack and ORFget is acceptable, it becomes prohibitive for ORFold. Furthermore, the sizes of the outputs are very large. In this section, we present alternatives to reduce the computational time and the size of the generated outputs.

3.2.1 FASTA and GFF files used in this example

- M_musculus.fna
- M_musculus.gff

(downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=mus+musculus)

- E_coli.fna
- E_coli.gff

(downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=e+coli)

- H_volcanii.fna
- H_volcanii.gff

(downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=haloferax+volcanii)

- D_melanogaster.fna
- D_melanogaster.gff

(downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=drosophila+melanogaster)

3.2.2 Annotation of ORFs of *Mus musculus* with ORFtrack

In order to reduce the execution time (around 64h hours on a single CPU), we recommend running ORFtrack on a cluster. The following command displays all the "seqid" contained in the first column of the input GFF file (usually chromosomes and contigs):

```
> ORFtrack -fna M_musculus.fna -gff M_musculus.gff --show-chr
```

The ORF annotation can be therefore distributed over multiple CPUs (i.e. one job per "seqid"), reducing substantially the computational time. That way, ORFtrack must be launched as many times as different "seqid" are indicated in the original GFF. Here, ORFtrack is launched on the chromosome NC_000067.7 with the following instruction:

```
> orftrack -fna M_musculus.fna -gff M_musculus.gff -chr
NC_000067.7
```

3.2.3 Extraction and writing of the ORFs and CDS of *Mus musculus* with ORFget

Definition of a minimal subset size to characterize the fold potential and structural properties of noncoding ORFs

Extracting all annotated ORFs with ORFget takes around 3 hours on a single CPU and generates a 7.5GB FASTA file containing up to 89×10^6 noncoding ORFs. Characterizing their fold potential and disorder and aggregation propensities with ORFold would take about 6 months on a single CPU. Consequently, we recommend running ORFold on a representative subset of noncoding ORFs. Indeed, a subset of 20000 ORFs is sufficient to estimate the fold potential and the disorder and aggregation propensities of the whole dataset of noncoding ORFs. The Kolmogorov Smirnov test p-value calculated for the comparison of the HCA score distribution obtained with a subset of 20000 randomly selected noncoding ORFs with that of the complete set of noncoding ORFs of *Drosophila melanogaster* is not significant. The same observations are made for the IUPred and Tango score distributions and hold also for other species such as *Haloferax volcanii* and *Escherichia coli*. Consequently, in the next section, ORFold will be applied to a set of 20000 randomly selected noncoding ORFs extracted from the complete set of mouse noncoding ORFs.

Extraction and writing of the amino acid sequences of a dataset of 20000 noncoding ORFs

The following instruction allows the extraction of a subset of 20000 noncoding ORFs (see Note 3 for more advanced examples).

```
> orfget -fna M_musculus.fna -gff mapping_orf_M_musculus.gff -
features_include nc -o M_musculus_noncoding -N 20000
```

Then, in order to compare the fold potential and the disorder and aggregation propensities of the noncoding ORFs of *Mus musculus* with those of the CDS, we reconstruct the amino acid sequences of all the isoforms annotated in the original GFF file.

```
> orfget M_musculus.fna -gff M_musculus.gff -features_include CDS
-o M_musculus_CDS
```

3.2.4 Characterization of the fold potential, and the structural properties of a set of 20000 noncoding ORFs along with those of *M. musculus* CDS

We execute ORFold on the small dataset of randomly selected noncoding ORFs and the complete set of mouse isoforms:

```
> orfold -fna M_musculus_noncoding.pfasta M_musculus_CDS.pfasta -
options HIT
```

ORFold provides us with two tables containing the fold potential and the disorder and aggregation propensities of the 20000 noncoding ORFs and the 92473 mouse isoforms (around 40 hours on a single CPU).

3.2.5 Comparison of the fold potential of the noncoding ORFs and the CDS calculated for different species

ORFplot can handle multiple datasets at the same time. Following the same protocol as the one used for the mouse, we also calculated the fold potential of a subset of 20000 noncoding ORFs and all CDS of *Haloferax volcanii*, *Escherichia coli*, and *Drosophila melanogaster*. We then present the HCA score distributions of all datasets on the same graph.

```
> orfplot -tab E_coli_CDS.tab H_volcanii_CDS.tab
D_melanogaster_CDS.tab M_musculus_CDS.tab -names “E. coli” “H.
volcanii” “D. melanogaster” “M. musculus”
```

```
> orfplot -tab E_coli_noncoding.tab H_volcanii_noncoding.tab
D_melanogaster_noncoding.tab mouse_noncoding.tab -names “E. coli”
“H. volcanii” “D. melanogaster” “M. musculus”
```

Figure 2.6 shows for the four species, the HCA score distributions of the corresponding CDS (Figure 2.6A) and noncoding ORFs (Figure 2.6B). Although the fold potential

distributions of the CDS display slight variations among the four species, the vast majority (more than 85%) exhibits intermediate HCA scores typical of the scores obtained for the globular proteins. This reflects that being foldable is a trait that has been strongly selected during evolution. However, the fold potential distribution of the noncoding ORFs calculated for *Haloferax volcanii* is clearly different from those of the other species. Indeed, the other species are mostly characterized by noncoding ORFs that, similarly to CDS, encode peptides predicted as foldable. Conversely, the noncoding ORFs of *Haloferax volcanii* are enriched in sequences with low HCA scores that are likely to encode disordered peptides. Whether this enrichment in hydrophilic sequences comes from the fact that this species lives in hypersaline environments is an exciting question that deserves further investigations.

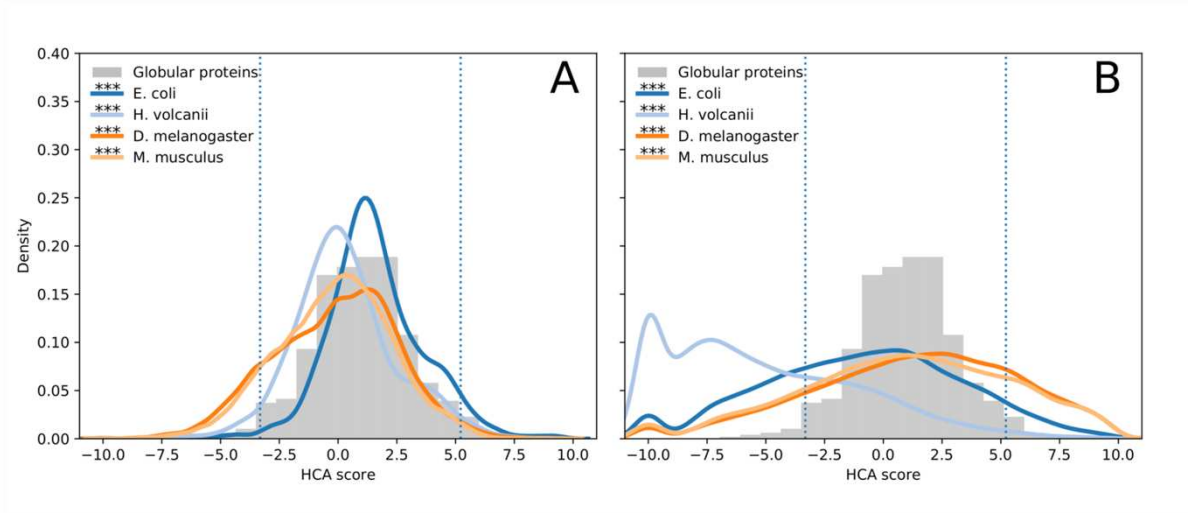


Figure 2.6: (A) Distribution of the HCA scores calculated for the CDS of *E. coli*, *H. volcanii*, *D. melanogaster* and *M. musculus* (dark blue, light blue, dark orange and light orange curves respectively). (B) Distribution of the HCA scores calculated for the noncoding ORFs of *E. coli*, *H. volcanii*, *D. melanogaster* and *M. musculus* (dark blue, light blue, dark orange and light orange curves respectively). The HCA score distribution of the set of globular proteins is presented with the grey histogram. Each distribution is compared with the one of the globular proteins set with a Kolmogorov Smirnov test. Asterisks on the plot denote level of significance: *** $< 1 \times 10^{-3}$.

3.3 Probing the fold potential of a set of mouse noncoding ORFs shown to be pervasively translated

Recently, Ruiz-Orera et al. (2018) revealed with ribosome profiling experiments, the translation of 721 ORFs in mouse lncRNAs (i.e. translated lncRNA-ORFs). They are not

conserved across the neighboring species nor subjected to selective pressure. The authors propose them as intermediates between noncoding ORFs and de novo genes (Ruiz-Orera et al. 2018). This prompts us to ask whether their corresponding peptides display specific structural properties compared to those of the ones encoded by ORFs in other lncRNAs (i.e. nontranslated lncRNA-ORFs). Therefore, in this section, we characterize their respective HCA score distributions along with those of the CDS and the subset of 20000 randomly selected noncoding ORFs defined in Subheading 3.2. The amino acid sequences of all translated products identified in Ruiz-Orera et al. (2018) (i.e. products coming from protein coding genes or noncoding regions) can be downloaded at: https://figshare.com/articles/dataset/Ruiz-Orera_et_al_2017_/4702375?file=10323906

We extracted the sequences of the 721 translated lncRNA-ORFs by searching the sequences containing either the "lncRNAa:translated:NC" or the "novel:translated:NC" pattern in their annotation. Then, 20000 nontranslated lncRNA-ORFs were extracted randomly from the GFF generated with ORFtrack in Subheading 3.2 with the following instruction:

```
> orfget -fna M_musculus.fna -gff mapping_orf_M_musculus.gff -
features_include nc_ovp_same-lncRNA -o M_musculus_nc_ovp_same-
lncRNA -N 20000
```

The amino acid sequences of the 721 translated lncRNA-ORFs and the 20000 nontranslated lncRNA- ORFs can be directly given as input to ORFold.

```
> orfold -fna M_musculus_nc_ovp_same-lncRNA.pfasta
M_musculus_translated_721_orfs.pfasta - options H
```

We subsequently plot the fold potentials of the four sets of ORFs with ORFplot:

```
> orfplot M_musculus_CDS.tab M_musculus_noncoding.tab
M_musculus_nc_ovp_same-lncRNA.tab
M_musculus_translated_721_orfs.tab -names "CDS" "Noncoding ORFs"
"Nontranslated lncRNA-ORFs" "Translated lncRNA-ORFs"
```

Figure 2.7 shows the HCA score distributions of the four sets of ORFs. If the nontranslated lncRNA-ORFs display similar HCA scores to noncoding ORFs (Kolmogorov Smirnov test, P = 0.46), the 721 translated lncRNA-ORFs exhibit a clearly different HCA value distribution from the three other datasets (Kolmogorov Smirnov test, P = 5.9e-06, 4.8e-06, and 2.4e-05 with nontranslated lncRNA- ORFs, noncoding ORFs, and CDS respectively). Although they are characterized by a majority of intermediate HCA score sequences expected to be foldable, they are clearly enriched in disorder prone sequences recalling the observation made by Wilson et al. (2017) that young proteins are more disordered than old ones. That said, it is interesting to note that, similarly to the two other noncoding ORF categories, the translated lncRNA-ORFs exhibit a majority of sequences that potentially encode peptides expected to be foldable. Further investigations are needed to determine whether their corresponding peptides fold to a well-defined and stable 3D structure or to a molten globule.

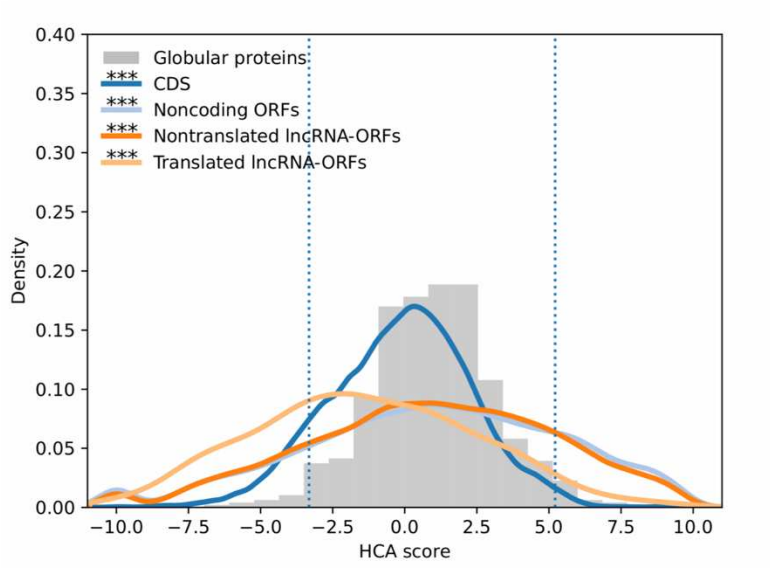


Figure 2.7: Distribution of the HCA scores calculated for the CDS, the 20000 noncoding ORFs, the 2000 nontranslated lncRNA-ORFs, and the 721 translated lncRNA-ORFs of *M. musculus* (dark blue, light blue, dark orange and light orange curves respectively). The HCA score distribution of the set of globular proteins is presented with the grey histogram. Each distribution is compared with the one of the globular proteins set with a Kolmogorov Smirnov test. Asterisks on the plot denote level of significance: *** < 1×10^{-3}.

4 Conclusion

Here, we presented three protocols that all aim at characterizing the fold potential and the structural properties of different sets of ORFs, including coding sequences, the ensemble or a representative subset of the noncoding ORFs of a genome, or a specific subset of sequences of interest. ORFtrack is very fast, annotating million ORFs in a few hours. In addition, it allows the user to deal with different levels of annotation and various combinations of selection patterns, thereby facilitating the definition of many ORF categories. ORFold can handle many inputs and enables the simultaneous visualization of the fold potential calculated for different datasets or the manual inspection of the fold potential or structural properties of all annotated ORFs of a genome with a genome viewer. In addition, ORFold can be used to probe the fold potential and the structural properties of any set of amino acid sequences without any genomic information including for instance, designed peptides or de novo peptides identified with mass spectrometry in different tissues or conditions. Finally, ORFmine opens up new applications in peptide discovery and characterization. In particular, recent studies have reported the existence of de novo peptides associated with human diseases (Barbosa et al. 2013; Lawrence et al. 2013; Yadav et al. 2014; Sendoel et al. 2017; von Bohlen et al. 2017; Wang et al. 2019; Yin et al. 2019). ORFtrack can be used to mine noncoding genomes for the identification of de novo peptides which are usually difficult to identify with mass spectrometry experiments (for example, peptides resulting from the translation of RNAs associated with diseases). On the other hand, ORFold provides valuable and complementary information with the characterization of their fold potential and structural properties.

5 Notes

1. Notice that the genomic features of a GFF3 file follow a specific hierarchy. For example, the feature “gene” has children (e.g. CDS, exons, tRNAs, rRNAs...). In addition, features of the same level can overlap with each other (e.g. a CDS and its corresponding exon). By default, the features “gene” and “exon” are not considered. ORFs that match with the feature “gene” will be annotated according to its children or

related features (mRNA, tRNA...). For example, ORFs overlapping tRNAs on the same strand necessarily overlap the parent genes as well, but for a more precise annotation, ORFtrack will annotate them as nc_ovp_same-tRNA instead of nc_ovp_same-gene. Finally, an ORF that matches with the feature "CDS", usually matches with the corresponding "exon" feature as well. However, the "exon" feature is not considered and the ORF will be annotated as c_CDS if it is in the same frame as the CDS, or as nc_(same/opp)_ovp-CDS if it is in another frame than the CDS.

2. Notice that the following instructions will lead to the same result.

```
> orfget -fna E_coli.fna -gff mapping_orf_E_coli.gff -
features_include nc_intergenic nc_ovp -o E_coli_noncoding
```

3. Notice that ORFget can extract a random subset of ORFs belonging to a specific category (e.g. extraction of 20000 noncoding ORFs overlapping lncRNAs on the same strand) as follows:

```
> orfget -fna M_musculus.fna -gff mapping_orf_M_musculus.gff -
features_include nc_ovp_same-lncRNA -o
M_musculus_nc_ORF_ovp_same-lnRNA -N 20000
```

6 References

Barbosa C, Peixeiro I, Romão L (2013) Gene expression regulation by upstream open reading frames and human disease. PLoS Genet 9:e1003529

Bartonek L, Braun D, Zagrovic B (2020) Frameshifting preserves key physicochemical properties of proteins. Proceedings of the National Academy of Sciences 117:5907–5912

Bitard-Feildel T, Callebaut I (2017) Exploring the dark foldable proteome by considering hydrophobic amino acids topology. Scientific reports 7:1–13

Bitard-Feildel T, Callebaut I (2018) HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences. bioRxiv 249995

Carvunis A-R, Rolland T, Wapinski I, et al (2012) Proto-genes and de novo gene birth. Nature 487:370–374

Chen J, Brunner A-D, Cogan JZ, et al (2020) Pervasive functional translation of noncanonical human open reading frames. Science 367:1140–1146

Deng Y, Bamigbade AT, Hammad MA, et al (2018) Identification of small ORF-encoded peptides in mouse serum. Biophysics reports 4:39–49

Dosztányi Z (2018) Prediction of protein disorder based on IUPred. Protein Science 27:331–340

Eguen T, Straub D, Graeff M, Wenkel S (2015) MicroProteins: small size–big impact. Trends in plant science 20:477–482

Erdős G, Dosztányi Z (2020) Analyzing Protein Disorder with IUPred2A. Current Protocols in Bioinformatics 70:e99

Faure G, Callebaut I (2013) Comprehensive repertoire of foldable regions within whole genomes. PLoS computational biology 9:

Faure G, Callebaut I (2013) Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. Bioinformatics 29:1726– 1733

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature biotechnology 22:1302–1306

Hobbs EC, Fontaine F, Yin X, Storz G (2011) An expanding universe of small proteins. Current opinion in microbiology 14:167–173

Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147:789–802

Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715–718

Lamiable A, Bitard-Feildel T, Rebehmed J, et al (2019) A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. Biochimie 167:68–80

Lawrence MS, Stojanov P, Polak P, et al (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218

Li J, Liu C (2019) Coding or noncoding, the converging concepts of RNAs. Frontiers in Genetics 10:496

Linding R, Schymkowitz J, Rousseau F, et al (2004) A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. Journal of molecular biology 342:345–353

Mészáros B, Erdős G, Dosztányi Z (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic acids research 46:W329– W337

Mészáros B, Simon I, Dosztányi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5:e1000376

Neme R, Amador C, Yildirim B, et al (2017) Random sequences are an abundant source of bioactive RNAs or peptides. Nature ecology & evolution 1:1–7

Prabakaran S, Hemberg M, Chauhan R, et al (2014) Quantitative profiling of peptides from RNAs classified as noncoding. Nat

Robinson JT, Thorvaldsdóttir H, Winckler W, et al (2011) Integrative genomics viewer. Nature biotechnology 29:24–26

Rousseau F, Schymkowitz J, Serrano L (2006) Protein aggregation and amyloidosis: confusion of the kinds? Current opinion in structural biology 16:118–126

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas J, et al (2018) Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat. Ecol. Evol. 2: 890–896. Nature ecology & evolution

Samayoa J, Yildiz FH, Karplus K (2011) Identification of prokaryotic small proteins using a comparative genomic approach. Bioinformatics 27:1765–1771

Schaefer C, Schlessinger A, Rost B (2010) Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics 26:625–631

Sendoel A, Dunn JG, Rodriguez EH, et al (2017) Translation from unconventional 5′ start sites drives tumour initiation. Nature 541:494–499

Slavoff SA, Mitchell AJ, Schwaid AG, et al (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nature chemical biology 9:59

Tretyachenko V, Vymětal J, Bednárová L, et al (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Scientific reports 7:1–9

von Bohlen AE, Böhm J, Pop R, et al (2017) A mutation creating an upstream initiation codon in the SOX 9 5′ UTR causes acampomelic campomelic dysplasia. Molecular genetics & genomic medicine 5:261–268

Wang S, Mao C, Liu S (2019) Peptides encoded by noncoding genes: challenges and perspectives. Signal Transduction and Targeted Therapy 4:1–12

Wilson BA, Foy SG, Neme R, Masel J (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature ecology & evolution 1:1–6

Yadav M, Jhunjhunwala S, Phung QT, et al (2014) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature 515:572–576

Yin X, Jing Y, Xu H (2019) Mining for missed sORF-encoded peptides. Expert Review of Proteomics 16:257–266

2.2 ORFribo

ORFribo corresponds to the third tool of ORFmine package (not still integrated) which aims at mapping correctly Ribo Seq data on the ORFs of a given genome being coding or not, thereby probing the translational activity of CDS but also of ORFs presumed as noncoding. As it has already been presented in the introduction, Ribosome Profiling is more informative method than the RNA Seq as it can identify specifically the codon under translation (P-site) and consequently the frame of the mRNA (out of the three possible) which is indeed translated. However, mapping Ribo Seq data on noncoding regions and detecting the frame of translation is one of the most complicated tasks of the Ribo Seq data analysis. The data need to be first calibrated on coding sequences (CDS), for which the translation frame is known, and then can be used for the detection of the frame that is translated in sequences unannotated and therefore presumed as noncoding.

2.2.1 Detection of the P-site through phasing of the reads on the transcriptome

Ideally, the size of the Ribo Seq reads is expected to be 28 nucleotides (corresponding to the length of the mRNA protected by the ribosome machinery during translation) and the first nucleotide of the codon under translation (P-site) is localized at the 14th position of the read. However, the experimental conditions (conformational changes of the ribosome machinery, alterations at the digestion time by the nuclease etc.) can generate reads of different sizes, thereby leading to the modification of the P-site's location on the reads, which may lead to misidentification of the correct P-site (Figure 2.8). One should note that, false localization of the P-site, inevitably leads to the wrong identification of the true frame under translation. Therefore, it is of crucial importance to identify the fragments size which enable us to correctly detect the location of the P-site and consequently of the translated frame. This procedure is called "phasing" and consists in separating the Ribo Seq reads into groups of different sizes (notably from 26 to 30 nucleotides) and mapping them independently on the transcriptome (CDS) of the organism. The advantage of using the CDS is that we know which frame among the three possible is expected to be translated. We can use this information in order to identify the reads' size for which the first nucleotide of a true codon under translation is localized at the 14th position for the

majority of the reads (Figure 2.8). By convention, we name the translated frame as "Frame 0" while the +1 and +2 frames are named as "Frame 1" and "Frame 2", respectively. Consequently, a read for which the 14th position indicates the first nucleotide of the true translated codon is considered as "in-frame" or a "Frame 0 read". In a similar way, a read for which the 14th position indicates the second or third nucleotide of the true translated codon is considered as a "Frame 1 read" or a "Frame 2 read", respectively. The phasing procedure aims at detecting the kmers or group of reads according to their size that maximizes the fraction of in-frame reads ("Frame 0") and minimizes the fraction of out-of-frame reads ("Frame 1" and "Frame 2"). Indeed, the 14th position of these kmers is expected to indicate correctly the codon under translation and subsequently the translated frame in the translated RNA.

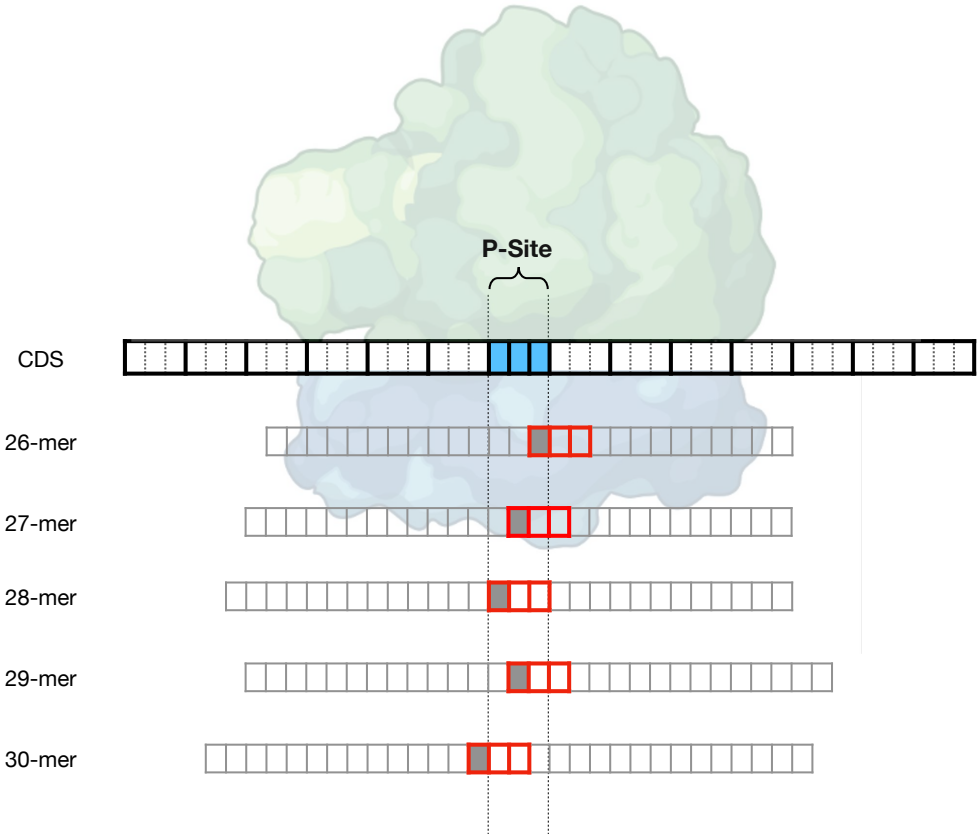


Figure 2.8. Schematical representation of the phasing procedure for Ribo Seq reads with different sizes aligned on the same region of the transcript. Every square corresponds to a single nucleotide while the codons on the CDS sequence are highlighted with thick black line. In blue is colored the codon under translation localized in the P-site of the ribosome. Ribo Seq reads of different sizes are aligned with the sequence of the translated mRNA and their 14th position is highlighted with grey filled square. Their theoretical codon under translation is highlighted with red thick line. As it can be observed, only the 28-mer identified correctly the true P-site and as a result this read is tagged as "Frame 0" because is in the

same frame with the coding sequence indeed translated. The 14th position of the 26- and 30-mer reads localize at third nucleotide of a codon of the translated frame and are tagged as "Frame 2 reads" while the 14th position of the 27- and 29-mer reads localize at the second nucleotide of a codon of the translated frame, being subsequently tagged as "Frame 1 reads". Only the "Frame 0" reads indicate codons indeed translated in the transcript. The ribosome figure was created with BioRender.com.

The phasing procedure is conducted by the ORFribo pipeline using the option "phase". The output of this step is a table of reads count per transcript together with their tagging as in-frame ("Frame 0") or out-of-frame reads ("Frame 1" and "Frame 2"). In addition, another table is generated, mapping the number of reads (and tagging them as in-frame or out-of-frame) per codon. This file permits us to study the periodicity of the reads which corresponds to the total count of reads tagged as "Frame 0", "Frame 1" and "Frame 2" per position on the transcripts. Due to the variant sizes of the transcripts, we can only visualize the periodicity for a few positions at the beginning and at the end of the transcripts.

In Figure 2.9 are presented the results of the phasing for one Ribo Seq experiment (accession number: SRR6398740) on the transcriptome of *S. cerevisiae*. Every horizontal line corresponds to reads of different sizes (from 26 to 30 nucleotides). Figure 2.9A represents the distribution of the reads tagged as "in-frame" ("Frame 0") and "out-of-frame" ("Frame 1" or "Frame 2") per transcript. It can be observed that ~70% of the 28-mers correspond to reads that localize the true P-site at their 14th position (tagged as "Frame 0") while the other two types of reads ("Frame 1" and "Frame 2") correspond to ~15%, each. As a result, the 28-mers seem to be the best phased reads and are the ones for which we can identify with higher confidence (~70% at this example) the true codon under translation and finally the translated frame. Figure 2.9B-C correspond to the periodicity of the reads at the beginning and the end of the transcripts, respectively. The periodicity plot informs us also about the quality of the data and therefore of the experiment. Again the 28-mers present a clear periodic signal with the "Frame 0" reads (in green) being overrepresented against the other two types of reads for every position of the transcripts. In addition, the periodicity plot informs us about the total number of reads per position which can be compared among the different read sizes. In this example, we can observe that the 28-mers not only are better phased and present more

periodic signals on the transcripts but also contain a more important number of reads than the 26- or the 27-mers.

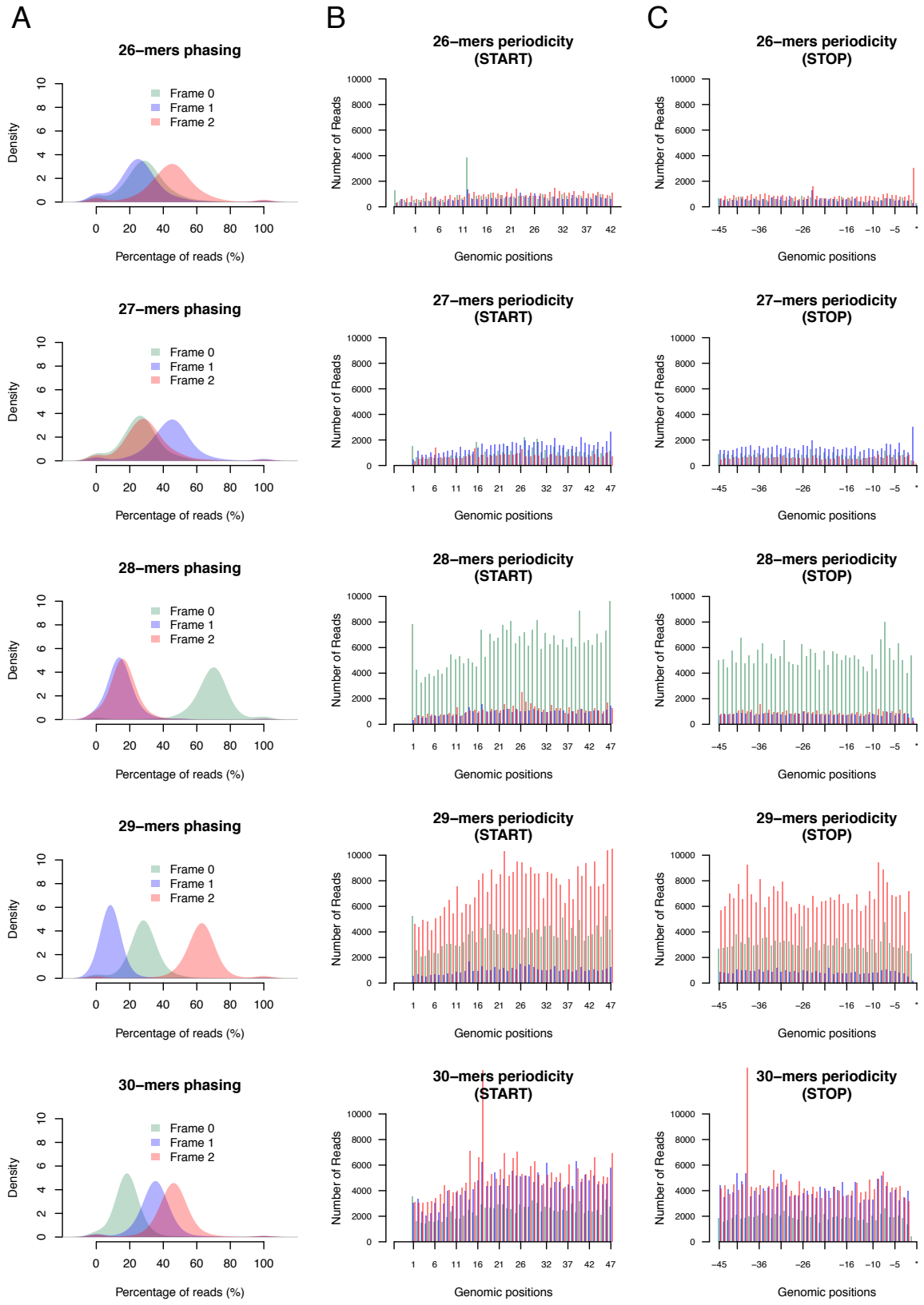


Figure 2.9. Example of the results of the phasing for one Ribo Seq experiment (accession number: SRR6398740) on the transcriptome of *S. cerevisiae.* **(A)** Distribution of the in-frame reads ("Frame 0") and out-of-frame reads ("Frame 1" and "Frame 2") for different read sizes (i.e., kmers). **(B)** Periodicity of the reads at the beginning of the transcripts **(C)** Periodicity of the reads at the end of the transcripts.

2.2.2 Detection of the frame under translation for intergenic mRNA

With the phasing procedure, we aim at detecting the kmer(s) that maximize(s) the fraction of in-frame reads and minimizes the fraction of the out-of-frame reads on the transcriptome, for which we know the frame expected to be translated. The phasing step is of crucial importance because (i) gives us information about the quality of the Ribo Seq data and therefore of the experiment itself and (ii) permits us to identify the read size for which we are highly confident about the detection of the frame that is indeed translated. Having good phasing (high fraction of in-frame reads) and good periodicity with the selected reads size, we can infer the translated frame of noncoding RNAs with mapped reads (for which by definition we do not know the translated frame). As a result, noncoding ORFs with in-frame mapped reads can be considered as truly translated (Figure 2.10).

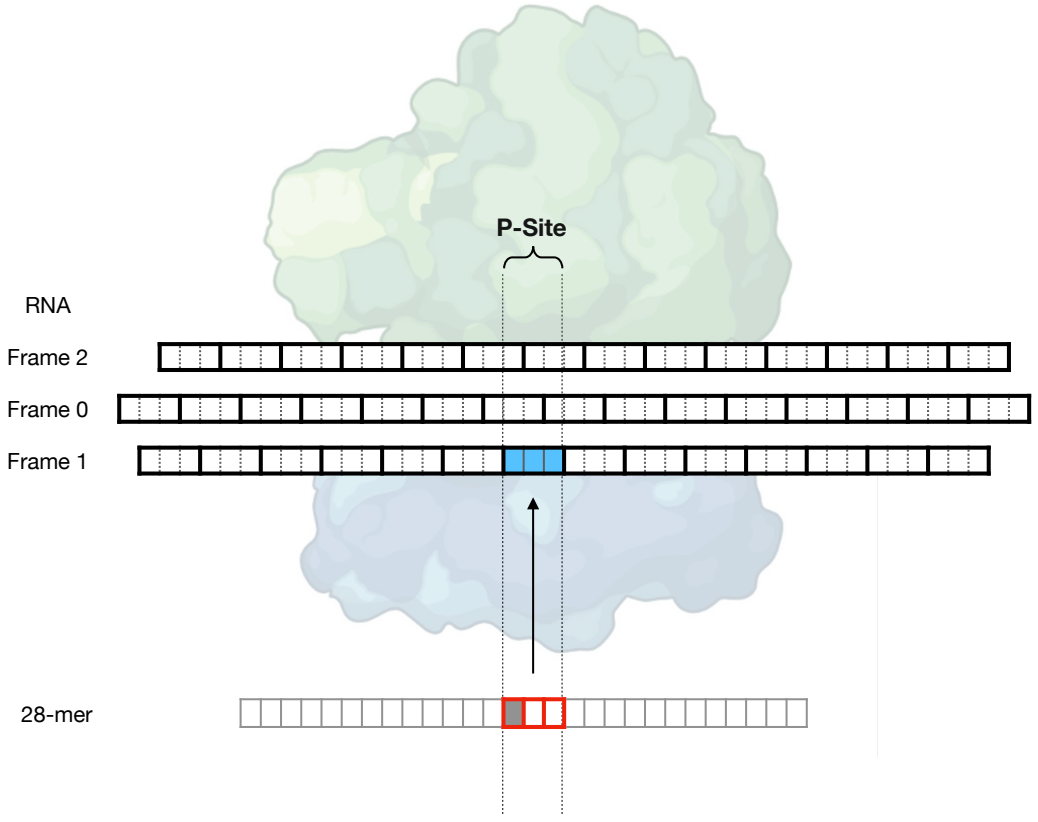


Figure 2.10. Schematical representation of the detection of the frame under translation for a noncoding RNA. Every square corresponds to a single nucleotide while the codons of the three possible Open Reading Frames of the mRNA sequence are highlighted with thick black line. In blue is colored the codon under translation localized in the P-site of the ribosome. The 28-mer Ribo Seq read is aligned with the sequence of the RNA and its 14th position is highlighted with grey filled square. Its theoretical codon under translation is highlighted with red thick line. Based on the phasing procedure (presented in Figures 2.8 and 2.9) the 28-mers were found to be correctly phased on the transcriptome and consequently are the ones expected to indicate the true P-site at their 14th position. As a result, 28-mers aligned on presumed noncoding regions can be used to identify which of the three possible Open Reading Frames of the RNA sequence was the one truly translated. The Frame 0, 1 and 2 correspond to the relative frames of the RNA molecule. In this example, the 28-mer aligned on the noncoding RNA specifically indicates the Frame 1 as the frame under translation and consequently determines the noncoding ORF (among the three overlapping) which was translated. The ribosome figure was created with BioRender.com.

2.2.3 Protocol for the mapping of the Ribo Seq reads

In Figure 2.11 is presented the pipeline for phasing and mapping Ribo Seq reads using the ORFribo tool in combination with other tools of ORFmine.

1. **Transcriptome extraction:** The phasing of the Ribo Seq reads is conducted on the transcriptome for which we know the correct frame of translation. With the genome (in fasta file) and its annotation (in gff file) and using the tool ORFget we can extract the nucleotide sequences of the transcriptome of the organism (in fasta file).
2. **Phasing the reads on the transcriptome:** Then with the transcriptome and the Ribo Seq raw reads (in fastq file) and using the ORFribo tool (option “phase”) we launch the phasing procedure. The initial step is to remove the adaptor sequence used during the sequencing procedure. The tool Cutadapt (v3.4) (Martin 2011) is used to remove the adaptor sequence from the 3’ end of the reads and filter the remaining nucleotide sequences based on their size (from 26 to 30 nucleotides). Then every set of read sizes is mapped on the transcriptome independently, and the counts of the reads per transcript are obtained with ORFribo (option “map”). The alignment of the reads on the transcriptome is conducted with the tool Bowtie (v1.3) (Langmead et al. 2009). We permit maximum 2 misaligned nucleotides per read and count only the reads that are aligned to a single region. Samtools (Li et

al. 2009a) is used in order to index the aligned reads and transform them into binary file decreasing the computational time. Then ORFribo detects the theoretical P-site on the mapped reads and tags every read as in-frame ("Frame 0"), if the read is in the same frame with a transcript, or out-of-frame ("Frame 1" or "Frame 2") in the opposite case. Then generates a table file containing the count of in-frame and out-of-frame reads per transcript.

3. **Finding the best kmer:** Based on the distribution of the in-frame reads per transcript, ORFribo can make an automated decision about the best phased kmer(s). The threshold of the fraction of in-frame reads, above which we consider a kmer to present a good phasing, is given by the user. The user can specify whether he prefers the threshold to be compared with the mean or the median value of the distribution of in-frame reads.
4. **Mapping the best phased kmer(s) on the IGORFs:** Once the best phased kmer(s) is/are detected, we can map the selected reads size(s) on the ensemble of the noncoding ORFs and detect with high confidence noncoding ORFs with in-frame reads, thus considered as translated. For that step we use ORFribo (option "map") and we provide as inputs the reads of the best phased kmer(s) (in fastq file), the genome sequence (in fasta file) and the total ORFs annotation (in gff file) as generated by ORFtrack. ORFribo generates a table containing the count of in-frame ("Frame 0") and out-of-frame ("Frame 1" and "Frame 2") reads for every ORF contained in the ORF annotation file.

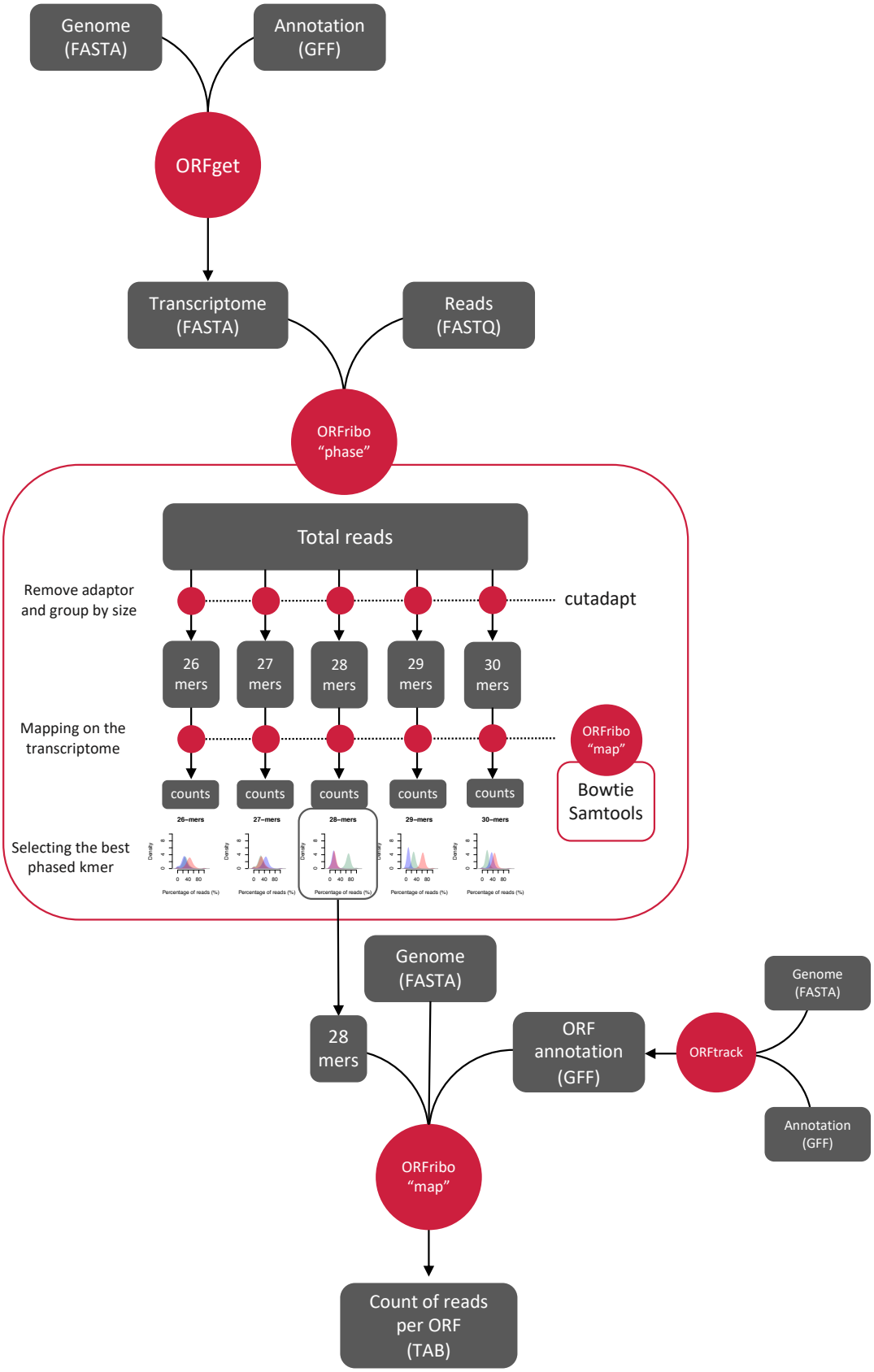

Genome (FASTA)
Annotation (GFF)
ORFget
Transcriptome (FASTA)
Reads (FASTQ)
ORFribo "phase"
Total reads
Remove adaptor and group by size
cutadapt
26 mers
27 mers
28 mers
29 mers
30 mers
Mapping on the transcriptome
ORFribo "map"
Bowtie
Samtools
counts
Selecting the best phased kmer
Genome (FASTA)
28 mers
ORF annotation (GFF)
ORFtrack
Genome (FASTA)
Annotation (GFF)
ORFribo "map"
Count of reads per ORF (TAB)

Figure 2.11. Pipeline of ORFribo. The inputs and outputs are represented with grey rectangles while the main scripts are shown with red circles. The mandatory inputs as well as their corresponding outputs are connected to their related scripts with black arrows. The pipeline starts with the extraction of the transcriptome's nucleotide sequences using the ORFget tool. Then follows the phasing of the Ribo Seq reads on the transcriptome. The adaptors are removed with cutadapt and the reads are organized based on their size (from 26 to 30 nucleotides). Every set of reads size is mapped independently on the transcriptome with ORFribo-"map" (using Bowtie and Samtools) and one table of read counts per transcript is generated for every read size group. The distribution of the reads in-frame per transcript is estimated and ORFribo decides which size(s) of reads is/are the best phased based on a threshold defined by the user. In this case, only the 28-mers passed the threshold and are mapped on the noncoding ORFs using again ORFribo-"map". This time we map the 28-mers on the total genome and count the mapped reads only for the ORFs indicated in the ORF annotation file as generated by ORFtrack.

The initial inputs for the ORFribo pipeline are (i) the genome sequence (in fasta file), (ii) the genome annotation (in gff file), (iii) the Ribo Seq reads raw data of a single experiment (in fastq file) and the ORF annotation (in gff file) generated by ORFtrack presented already. The final output is a table format file with the count of reads (in-frame and out-of-frame) for every ORF indicated in the ORF annotation file. The automatization of the Ribo Seq data mapping on the noncoding genome permits us to apply this analysis on multiple datasets. Due to the noncoding ORFs' low signal of translation, combining the results of multiple different experiments is expected to increase the probability of detecting noncoding ORFs under translation. In addition, the systematic detection of translation signal for some noncoding ORFs permits us to discriminate between truly translated noncoding ORFs and ribosomes which are simply scanning the RNA.

3 Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution

In this section I applied the methods presented in Section 2 in order to detect all the IGORFs of *S. cerevisiae* (with ORFtrack) and estimate the fold potential diversity (i.e., propensity for disorder, folded state, or aggregation) together with other sequence and structural properties of the peptides encoded by them (with ORFold). This permitted me to explore the foldability diversity encoded by ORFs hosted in the noncoding genome of the yeast and compare it with the one observed in proteomes. Overall, it permitted me to estimate the potential of the noncoding genome to produce novel structural bricks which can either serve as starting points for the birth of novel genes or be integrated into pre-existing proteins.

Then, using comparative genomics and ancestral reconstruction I systematically reconstructed the ancestral noncoding sequences that gave rise to 70 known de novo genes of *S. cerevisiae* and characterized their sequence and structural properties. This permitted me to compare the fold potential of de novo proteins with the one of the ensemble of IGORFs in order to understand whether IGORFs that gave birth to novel genes display specific sequence and structural properties. Finally, analyzing ribosome profiling data of five independent experiments (with ORFribo), I identified IGORFs with a strong translation signal in order to investigate the sequence and structural properties of candidate IGORFS that could potentially give birth to future novel genes.

All the results of this section are presented in a research article entitled "*Intergenic ORFs as elementary structural modules of* de novo *gene birth and protein evolution*" and which has been published at the Genome Research peer-reviewed journal (Papadopoulos et al. 2021).

Title

Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution

Authors

Chris Papadopoulos[1], Isabelle Callebaut[2], Jean-Christophe Gelly[3,4,5], Isabelle Hatin[1], Olivier Namy[1], Maxime Renard[1], Olivier Lespinet[1], Anne Lopes[1]

Affiliations

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.
[2] Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005, Paris, France
[3] Université de Paris, Biologie Intégrée du Globule Rouge, UMR_S1134, BIGR, INSERM, F-75015, Paris, France.
[4] Laboratoire d'Excellence GR-Ex, Paris, France
[5] Institut National de la Transfusion Sanguine, F-75015, Paris, France

Keywords

de novo genes, noncoding genome, foldability, genome evolution, protein evolution, protein bricks

Abstract

The noncoding genome plays an important role in de novo gene birth and in the emergence of genetic novelty. Nevertheless, how noncoding sequences' properties could promote the birth of novel genes and shape the evolution and the structural diversity of proteins remains unclear. Therefore, by combining different bioinformatic approaches, we characterized the fold potential diversity of the amino acid sequences encoded by all intergenic ORFs (Open Reading Frames) of *S. cerevisiae* with the aim of (i) exploring whether the structural states' diversity of proteomes is already present in noncoding sequences, and (ii) estimating the potential of the noncoding genome to produce novel protein bricks that could either give rise to novel genes or be integrated into pre-existing proteins, thus participating in protein structure diversity and evolution. We showed that amino acid sequences encoded by most yeast intergenic ORFs contain the elementary building blocks of protein structures. Moreover, they encompass the large structural state diversity of canonical proteins with the majority predicted as foldable. Then, we investigated the early stages of de novo gene birth by reconstructing the ancestral sequences of 70 yeast de novo genes and characterized the sequence and structural properties of intergenic ORFs with a strong translation signal. This enabled us to highlight sequence and structural factors determining de novo gene emergence. Finally, we showed a strong correlation between the fold potential of de novo proteins and the one of their ancestral amino acid sequences, reflecting the relationship between the noncoding genome and the protein structure universe.

Introduction

Comparative genomics have revealed the existence of an important amount of taxonomically restricted genes and more specifically of orphan genes in various eukaryotic genomes (Tautz and Domazet-Lošo 2011; Wissler et al. 2012; Van Oss and Carvunis 2019; Vakirlis et al. 2020b). These genes lack detectable homologs in outgroup species and can constitute up to 30% of a genome's genes. They can derive from clearly distinct mechanisms, including the well-known mechanisms of duplication or horizontal gene transfer followed by fast divergence (Kaessmann 2010; Tautz and Domazet-Lošo 2011; Schlötterer 2015; Van Oss and Carvunis 2019). However, de novo emergence from noncoding regions has now been proven to be an undeniable additional mechanism and studies reporting evidence of de novo gene birth are published every year, thereby giving a new role to noncoding regions in the creation of genetic novelty (Knowles and McLysaght 2009; Wu et al. 2011; Tautz and Domazet-Lošo 2011; Murphy and McLysaght 2012; Zhao et al. 2014; Schlötterer 2015; Li et al. 2016; Vakirlis et al. 2018; Zhang et al. 2019a; Vakirlis et al. 2020b; Heames et al. 2020; Blevins et al. 2021). Nevertheless, how noncoding sequences can code for a functional product and consequently give rise to novel genes remains unclear. Indeed, function is intimately related to protein structure and more generally to protein structural properties. All proteomes are characterized by a large diversity of structural states. The structural properties of a protein result from its composition in hydrophobic and hydrophilic residues. Highly disordered proteins display a high hydrophilic residue content. Membrane proteins which fold in lipidic environments, but aggregate in solution, are enriched in hydrophobic residues. Finally, foldable proteins are characterized by a subtle equilibrium of hydrophobic and hydrophilic residues (Bresler and Talmud 1944). The latter are arranged together into specific patterns that dictate the formation of the secondary structures and the outcoming fold. However, contrarily to coding sequences (CDS), the nucleotides of noncoding ones are expected to be distributed randomly along the DNA, thereby resulting in different amino acid compositions from CDS. If and how these amino acid compositions can account for the structural states observed in proteomes is a crucial question to understand the relationship, if any, between the noncoding genome and the protein structure universe. So far, different models of de novo gene emergence have been proposed (Carvunis et al. 2012; Schlötterer 2015; Wilson et al. 2017). The

"preadaptation" model proposes an "all or nothing transition to functionality" where only sequences pre-adapted not to be harmful (i.e. with enough disorder not to be subjected to aggregation), will give rise to gene birth (Wilson et al. 2017). This model is supported by the observation that young genes and de novo protein domains display a higher disorder propensity than old genes (Ekman and Elofsson 2010; Bitard-Feildel et al. 2015; Schmitz et al. 2018; Foy et al. 2019). In contrast, the proto-gene model proposes an evolutionary continuum ranging from nongenic sequences to genes (Carvunis et al. 2012). Here, genes evolve de novo through transitory proto-genes that result from pervasive expression of nongenic sequences and proto-genes are expected to exhibit features intermediate between non-gene and genes. In this study, the authors reported that in yeast, young genes are less prone to disorder. Recently, Vakirlis et al. (2020a) proposed a TM-first model where the membrane environment provides a safe niche for transmembrane (TM) adaptive emerging peptides which can further evolve toward more soluble peptides. These adaptive peptides have been identified with overexpression which, according to the authors, may not be reached outside the laboratory. Whether such peptides, though beneficial in the experimental conditions, would be produced and be beneficial in "natural" conditions, deserves further investigation.

Overall, all these studies attribute to the fold potential of noncoding ORFs (including the propensities for disorder, folded state, and aggregation) an important role in the emergence of genetic novelty. However, several questions remain open. First, if the sequence and structural properties of de novo genes have been largely investigated in specific species, the raw material for de novo gene birth and the early stages preceding the fixation of the beneficial ORFs are to be further characterized (Schmitz et al. 2018). Second, if the role of the noncoding genome in de novo gene birth has been largely investigated, its role in protein evolution and structural diversity is to be further characterized as well. Indeed, de novo domains may emerge from noncoding regions through ORF extension or exonization of introns (Bornberg-Bauer and Alba 2013; Bornberg-Bauer et al. 2015). On the other hand, we can assume that protein-coding genes, whatever their evolutionary history, have had a noncoding ancestral origin (Nielly-Thibault and Landry 2019). Whether the noncoding ORFs which gave rise to novel genes can account for the structural diversity of proteomes or whether this structural diversity evolved from ancestral genes which all displayed similar structural properties

(i.e. disordered, foldable or TM-prone) is a crucial question to better understand the role, if any, of noncoding sequences in the protein structure universe.

Here, we characterized the diversity of the fold potential encoded in all intergenic ORFs (IGORFs) of *S. cerevisiae* with the aim of (i) exploring whether the large diversity of structural states observed in proteomes is already present in noncoding sequences, and (ii) studying the potential of the noncoding genome to produce novel protein bricks that could give birth to novel genes or be integrated into pre-existing proteins. Then, we investigated the sequence and structural factors determining de novo gene emergence by (i) characterizing the early stages of de novo gene birth through the reconstruction of 70 yeast de novo genes' ancestral sequences and (ii) characterizing the sequence and structural properties of IGORFs with a strong translation signal through ribosome profiling experiments.

Results

We extracted 105041 IGORFs of at least 60 nucleotides in *S. cerevisiae* (Methods). We probed their fold potential with the Hydrophobic Cluster analysis (HCA) approach (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2017; Bitard-Feildel et al. 2018; Bitard-Feildel and Callebaut 2018a) and compared it with the one of the 6669 CDS of *S. cerevisiae*. HCA highlights from the sole information of a single amino acid sequence, the building blocks of protein folds that constitute signatures of folded domains. They consist of clusters of strong hydrophobic amino acids that have been shown to be associated with regular secondary structures (Bitard-Feildel and Callebaut 2017; Bitard-Feildel et al. 2018; Lamiable et al. 2019) (Supplemental Fig. S3.1). These clusters are connected by linkers corresponding to loops or disordered regions. The combination of hydrophobic clusters and linkers in a sequence determines its fold potential. The latter can be appreciated in a quantitative way through the calculation of a foldability score (HCA score) which covers all the fold potential diversity of proteins.

IGORFs contain elementary building blocks of proteins

We first investigated the structural and sequence properties of proteins encoded by CDS and IGORFs (Figure 3.1; Supplemental Tables S3.1-3.4). CDS are longer than IGORFs and contain more HCA clusters (Mann-Whitney *U* test, $P < 2.2 \times 10^{-16}$ for both observations) (Figure. 3.1A,B). The HCA clusters of CDS and IGORFs display similar sizes of about 11 residues (Mann-Whitney *U* test, $P = 1 \times 10^{-1}$) (Figure 3.1C) and 96.9% of IGORFs harbor at least one HCA cluster. This result shows that the elementary building blocks of proteins are widespread in noncoding sequences. In contrast, CDS are enriched in long linkers reflecting long flexible regions (6.3 and 11.5 residues for IGORFs and CDS on average respectively, Mann-Whitney *U* test, $P = 2.6 \times 10^{-11}$) (Figure 3.1D). As a control, we generated scrambled intergenic sequences (Methods). The resulting random IGORFs behave similarly to real IGORFs for most properties, while being slightly shorter (Mann-Whitney *U* test, $P = 3 \times 10^{-3}$) (Supplemental Fig. S3.2). Whether the enrichment in long ORFs observed for real IGORFs results from high GC content genomic regions (STOP codons are AT-rich) is to be further investigated.

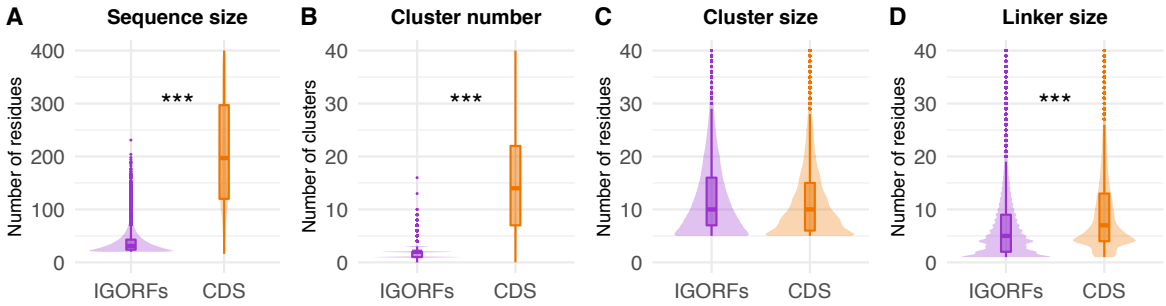


Figure 3.1: Plots of the distributions of sequence and HCA-based structural properties of IGORFs and CDS. **(A)** sequence size **(B)** number of HCA clusters per sequence **(C)** size of HCA clusters **(D)** size of linkers. The p-values were computed with the Mann-Whitney *U* test (one-sided for (A), (B), (D), and two-sided for (C)). Asterisks denote level of significance: ***$p < 1 \times 10^{-3}$, see Supplemental Tables S3.1-3.4 for detailed p-values.

CDS are enriched in polar and charged residues

If hydrophobic clusters of CDS and IGORFs display similar sizes, they may not have the same amino acid composition. Therefore, for each amino acid, we calculated its propensity for being in HCA clusters of CDS over HCA clusters of IGORFs. CDS HCA clusters are clearly enriched in polar and charged residues compared to those of IGORFs (Supplemental Fig. S3.3A). The same tendency is observed for CDS linkers (Supplemental Fig. S3.3B). Moreover, negatively charged residues are over-represented compared to positively charged ones in both HCA clusters and linkers of CDS. In fact, it has been shown that the charge distribution of a protein has an impact on its diffusion in the cytosol where positively charged proteins get caught in nonspecific interactions with the abundant negatively charged ribosomes (Schavemaker et al. 2017). We show that the frequency of negatively charged residues of the yeast cytoplasmic proteins is strongly correlated with the proteins' abundance (Spearman's correlation coefficient: Rho = 0.44, $P < 2 \times 10^{-16}$) suggesting that the crowded cellular environment has shaped the charge distribution of abundant proteins (Supplemental Fig. S3.4). This result recalls the observation made in previous studies showing that the frequency of “sticky” amino acids on the surface of globular proteins or in disordered proteins decreases as the protein cellular concentration increases (Levy et al. 2012; Macossay-Castillo et al. 2019). Finally, CDS tend to be enriched in ancient amino acids and codons and depleted in recent ones (Supplemental Fig. S3.5). As observed in other studies (Trifonov 1987; Brooks and Fresco 2003), yeast CDS are particularly enriched in GNN codons which include those coding for negatively charged amino acids. Whether this enrichment is unrelated to codon age and simply results from amino acid content constraints, whether CDS favor the usage of old codons for ignored reasons, or whether this observation results from a combination of both remains unclear.

IGORFs encode for peptides that display a wide diversity of fold potential including a substantial amount of foldable peptides

We next used the HCA score in order to assess the fold potential of the peptides encoded by IGORFs. As a reference, we calculated the HCA scores for three sequence datasets consisting of 731 disordered regions, 559 globular proteins and 1269 TM regions

extracted from transmembrane proteins, thereby expected to form aggregates in solution while being able to fold in lipidic environments (Methods) (Figure 3.2A; Supplemental Fig. S3.6). Based on their HCA scores, we defined three categories of fold potentials (i.e. disorder prone, foldable, or aggregation prone in solution). Here, we define as foldable, proteins that are able to fold into a compact and well-defined 3D structure or partially into an ordered structure in which the secondary structures are however present. Figure 3.2B shows that CDS and IGORFs belonging to the low HCA score category are indeed presumed to be disordered and display low propensity for aggregation. Comparable but small proportions of CDS and IGORFs fall into this group (4.9% and 7.7% respectively) indicating that most coding but also noncoding sequences are not highly prone to disorder in line with Tretyachenko et al. (2017). The high HCA score category corresponds to aggregation-prone sequences with low disorder propensity. CDS falling into this category are highly hydrophobic (Supplemental Table 3.5) with 81% of them annotated as uncharacterized according to UniProt (The UniProt Consortium 2019) and 60% predicted as containing at least one TM domain (Methods). Finally, the intermediate category gathers sequences which have a high potential for being completely or partially folded in solution as shown by their intermediate HCA scores comparable to those of globular proteins. Most CDS (91.4%) and a majority of IGORFs (66.6%) fall into this category. Both are characterized by intermediate aggregation and disorder propensities, although IGORFs display a wider range of aggregation propensities (Figure 3.2B). The fact that these CDS, though predicted as foldable, exhibit a certain propensity for aggregation, is in line with several studies which report a high aggregation propensity of proteomes across all kingdoms of life (Greenwald and Riek 2012; Langenberg et al. 2020). This observation has been explained as the side effect of the requirement of a hydrophobic core to form globular structures (Rousseau et al. 2006b; Ganesan et al. 2016; Langenberg et al. 2020). In particular, Langenberg et al. (2020), show a strong relationship between protein stability and aggregation propensity with aggregation prone regions mostly buried into the protein and providing stability to the resulting fold. Like for CDS, these regions, under the hydrophobic effect, may facilitate the stabilization of the IGORF encoded peptide structure. Whether peptides encoded by IGORFs in the intermediate category fold into a specific 3D structure, a partially ordered structure or a "rudimentary fold" which stabilizes itself through oligomerization like the Bsc4 de novo protein (Bungard et al. 2017), deserves further investigation. Finally, the proportions of

sequences in the different fold potential categories are different between IGORFs and CDS, with CDS mostly falling into the intermediate HCA score category reflecting that being foldable is a trait which has been strongly selected by evolution. In contrast, IGORFs cover a wide range of fold potentials that is also observed in random IGORFs (4.4%, 61.7% and 33.9% of sequences in the low, intermediate and high HCA score categories) showing that randomly, a wide range of fold potentials including a majority of foldable IGORFs can be expected. Overall, it is questionable whether de novo genes mainly originate from IGORFs encoding foldable peptides or from IGORFs whose corresponding peptides subsequently evolved toward foldable peptides regardless of their initial fold potential.

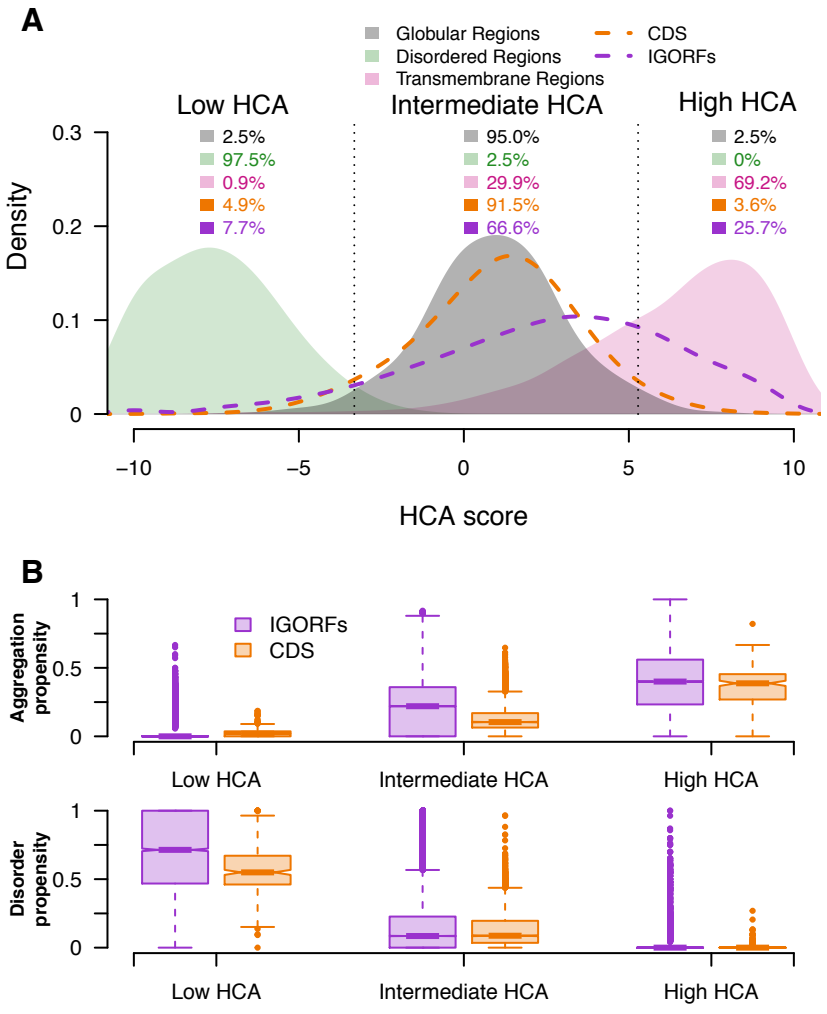


Figure 3.2. IGORFs encompass the large spectrum of fold potential of canonical proteins. **(A)** Distribution of the HCA scores for the three reference datasets (i.e. disordered regions, globular domains, and transmembrane regions - green, black and pink curves respectively) along with those for the CDS (orange curve) and IGORFs (purple curve). There is a clear distinction between the distributions of HCA scores

calculated for the three reference datasets. (Two-sided Kolmogorov-Smirnov test, $P < 2 \times 10^{-16}$ for all comparisons). Dotted black lines delineate the boundaries of the low, intermediate, and high HCA score categories reflecting the three categories of fold potential (i.e. disorder prone, foldable, or aggregation-prone in solution). The boundaries are defined so that 95% of globular domains fall into the intermediate HCA score category whereas the low and high HCA score categories include all sequences with HCA values that are lower or higher than those of 97.5% of globular domains respectively. High HCA scores reflect sequences with high densities in HCA clusters that are likely to form aggregates in solution. Low HCA scores indicate sequences with high propensities for disorder, while intermediate scores correspond to globular proteins characterized by an equilibrium of hydrophobic and hydrophilic residues (Methods). The percentages of sequences in each category are given for all datasets. Raw data distributions are presented in Supplemental Fig. S3.6. **(B)** Aggregation and disorder propensities calculated with TANGO and IUPred respectively are given for CDS and IGORFs of each foldability HCA score category.

From IGORFs to de novo genes

Therefore, we traced back the evolutionary events preceding the emergence of 70 de novo genes identified in S. cerevisiae by reconstructing their ancestral IGORFs (ancIGORFs) in order to compare the foldability potential of the peptides encoded by IGORFs that gave birth to de novo genes with the one of the peptides encoded by all other IGORFs and to characterize the steps preceding the emergence of a novel gene (Methods; Supplemental Fig. S3.7; Supplemental Table 3.6). Supplemental Fig. S3.8 shows the example of the YOR333C de novo gene which emerged in the lineage of S. cerevisiae. The corresponding noncoding region in the ancestors preceding its emergence consists of two IGORFs separated by a STOP codon. The fusion of the two consecutive IGORFs was triggered by two nucleotide substitutions which occurred specifically in the S. cerevisiae lineage and led respectively to the appearance of a start codon (mutation of Isoleucine into Methionine through an A/G substitution) and the mutation of the STOP codon into a Tyrosine through a G/C substitution. Overall, the 70 de novo genes emerged from a total of 167 ancIGORFs. A minority of de novo genes (16 cases) emerged from a single ancIGORF which covers almost all their sequence (95% of coverage between the ancestral IGORF and the resulting de novo gene) (i.e. single-ancIGORF de novo genes), while, the majority (54 cases) results from the combination of multiple ancIGORFs (2.8 on average) through insertion/deletion (indel) events leading to frameshifts in the original sequence and/or STOP codon mutations as observed with the example of YOR333C (i.e. multiple-ancIGORF de novo genes). In line with the findings of Zhang et al.

(2019), indels are two times more frequent than STOP codon mutations (64/33). Moreover, the multiple-ancIGORF de novo genes exhibit sequence sizes similar to those of the single-ancIGORF ones although the ancestral IGORFs they originate from are shorter than those that led to single-ancIGORFs de novo genes (Supplemental Fig. S3.9).

Figure 3.3 shows the HCA scores of the proteins encoded by the 70 de novo genes (i.e. de novo proteins) and of the peptides encoded by their corresponding ancIGORFs. The majority of de novo proteins (78%) are predicted as foldable, whereas peptides encoded by ancIGORFs display a larger range of HCA scores. However, ancIGORFs are not IGORF-like, being enriched in sequences encoding foldable peptides (75.4% and 66.6% for ancIGORFs and IGORFs respectively - one proportion z-test, $P = 9.5 \times 10^{-3}$) and depleted in sequences encoding aggregation prone ones (18.6% and 25.7% for ancIGORFs and IGORFs respectively, one proportion z test, $P = 2.1 \times 10^{-2}$).

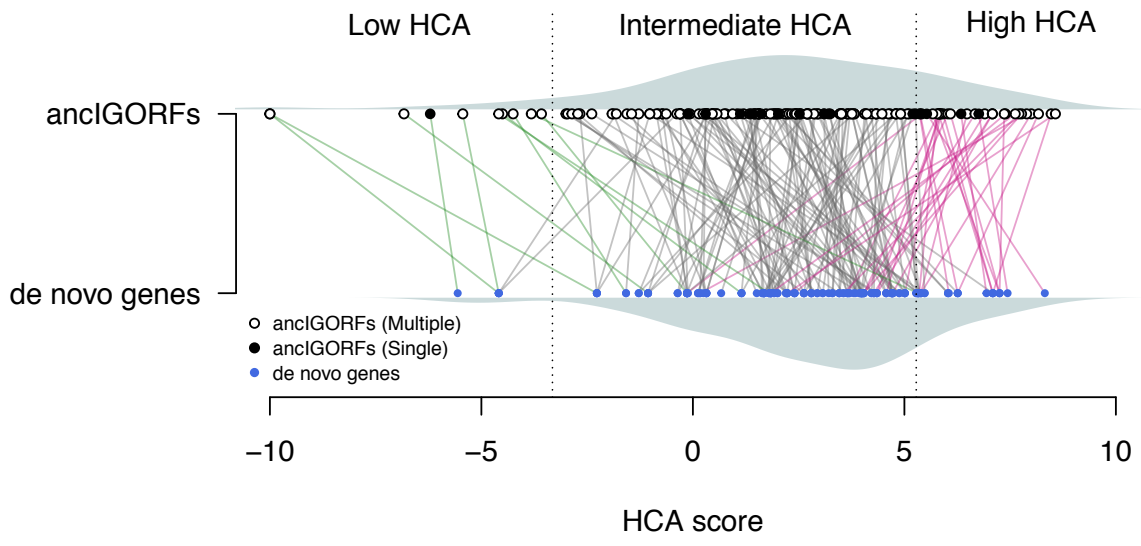


Figure 3.3. From ancIGORFs to de novo genes. Plot of the HCA score of each ancIGORF (black and white points for single and multiple ancIGORFs respectively) along with its corresponding de novo gene (blue points). Each de novo gene is connected to its parent ancIGORF(s) with a colored line. One should notice that a de novo gene is connected to several IGORFs when it results from the combination of different ancestral IGORFs (i.e. multiple-ancIGORF de novo genes). Green lines indicate cases where a de novo gene is connected to a low HCA score ancIGORF, while grey and pink lines indicate connections with an intermediate and a high HCA score ancIGORFs, respectively. The HCA score densities of de novo genes and ancIGORFs are shown in grey (bottom and top of the graph respectively).

Impact of indels and STOP codon mutations on the fold potential of a de novo protein

The overall relationship between the HCA scores of peptides encoded by ancIGORFs and their corresponding de novo proteins is characterized by a funnel shape revealing that most de novo proteins are foldable regardless of the fold potential of the peptides encoded by their IGORF parents (Figure 3.3). Two hypotheses can explain this observation: (i) this funnel mostly results from the amino acid substitutions which have occurred since the fixation of the ancIGORF(s) and which led to an increase in foldability of the resulting de novo genes, (ii) this funnel results from the fact that combining at least one IGORF encoding a foldable peptide with IGORFs encoding peptides with different fold potentials, leads to a foldable product. Figure 3.4A shows that de novo genes display amino acid frequencies similar to those of ancIGORFs (Supplemental Table S3.5). This result shows that the mutations which occurred since the fixation of the ancIGORF did not change the overall amino acid composition of the resulting de novo genes and thus, cannot explain the funnel shape observed in Figure 3.3. We then reasoned that since the divergence of the last common ancestor predating the emergence of de novo genes, single and multiple-ancIGORF de novo genes were subjected to similar amino acid mutation rates (average sequence identity between ancIGORFs and their corresponding de novo genes: 83% and 80% respectively), while the multiple-ancIGORF ones (which by definition result from the combination of several IGORFs) have also undergone indels and/or STOP codon mutations. This enabled us to quantify the impact of these different mutational events on the fold potential of the outcoming de novo proteins by calculating the correlation between the HCA score of each de novo protein and the peptides encoded by its corresponding ancIGORF(s). Figure 3.4B shows that single-ancIGORF de novo proteins display a clear correlation of HCA scores with those of the peptides encoded by their corresponding ancIGORFs (Spearman's correlation coefficient: Rho = 0.87, $P < 1.2 \times 10^{-5}$). This reveals that the amino acid mutations which occurred between the ancestor and the de novo protein did not affect the fold potential of the ancestral sequences suggesting that the structural properties of the peptides encoded by the single-ancIGORFs were retained in the resulting de novo proteins. In contrast, the correlation is weaker for multiple-ancIGORF de novo proteins (Spearman's correlation coefficient: Rho = 0.47, $P < 1.2 \times 10^{-9}$). This can be attributed to the fact that 81% (44/54) of the multiple-ancIGORF de novo proteins are predicted as foldable (white dots included in the pink

squares in Figure 3.4B) while being associated with ancIGORFs of different foldability potentials. All foldable de novo genes include at least one foldable ancestral peptide suggesting that in these cases, combining disordered or aggregation-prone peptides with a foldable one, has led to a foldable de novo protein as well. Supplemental Fig. S3.7E shows the example of the de novo gene YLL020C which results from the combination through an indel event, of a long foldable ancIGORF with a short IGORF predicted as aggregation prone. The resulting de novo gene is also predicted as foldable. Whether the foldable IGORF was the first to be selected and whether selection has only retained the combinations of IGORFs that do not affect the foldability of the preexisting selected product are exciting questions that deserve further investigation.

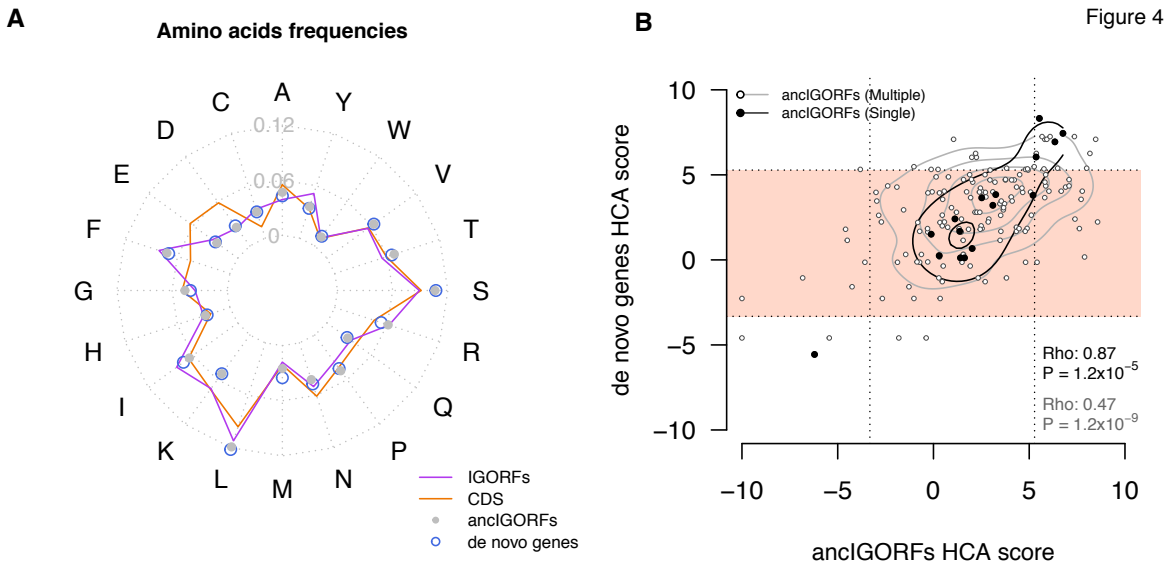


Figure 3.4. (A) Radar plot reflecting the 20 amino acid frequencies of IGORFs, ancIGORFs, de novo genes, and CDS. **(B)** Plot of the HCA score of each de novo gene with those of its parent ancIGORF(s). The fold potential of a single ancIGORF de novo gene is mostly determined by the one of its parent ancIGORFs while the combination of several ancIGORFs through frameshift events and STOP codon mutations leads most of the time to a foldable product. Single and multiple ancIGORF de novo genes are represented by black and white points respectively. Spearman's correlation coefficients of the relationships between single and multiple ancIGORF de novo genes' HCA scores versus the score of their parent ancIGORF(s), and the corresponding p-values are indicated on the plot. The contour lines mark the percentiles of the density function range in black and grey for single and multiple ancIGORF de novo genes respectively. The light pink region indicates de novo genes encoding proteins predicted as foldable.

Translation of IGORFs

Next, we performed ribosome profiling experiments on *S. cerevisiae* (strain BY4742) and used three additional ribosome profiling datasets to define two types of translated IGORFs (Radhakrishnan et al. 2016; Thiaville et al. 2016) (Methods). The former corresponds to IGORFs that are occasionally translated with a weak translation signal (at least 10 reads in one experiment) (Methods). The latter corresponds to IGORFs with a strong translation signal (more than 30 reads in at least two experiments) and whose translation is strongly favored over the overlapping IGORFs in the other phases (i.e. highly translated IGORFs) (Methods). We identified 1235 occasionally translated IGORFs and 31 highly translated ones. Figure 3.5 and Supplemental Fig. S3.10 show the frequencies of the first translated codons and amino acids respectively. For both highly and occasionally translated IGORFs, the first translated codon is enriched in AUG compared to all the other translated positions (one proportion z-test, both p-values $< 1 \times 10^{-16}$). The enrichment in AUG is clearly stronger for highly translated IGORFs, while the first translated codons of occasionally translated IGORFs are also enriched in the NUG near-cognate codons reported as alternative start codons (one proportion z-tests, all p-values $< 2 \times 10^{-2}$) (Ingolia et al. 2011; Cuevas et al. 2021). Nevertheless, due to the low read coverage of IGORFs, we cannot ensure that the first codon with a read is the first to be translated, though the enrichments in AUG or near-cognate codons support this assumption. In addition, the frequencies of the three STOP codons are comparable between all ORF categories (chi-2 tests between all pairs, p-values $> 5 \times 10^{-2}$) (Supplemental Table 3.7) with a systematic higher frequency of UAA. *S. cerevisiae* genome is AT-rich and the clear enrichment in UAA in all ORF categories including IGORFs, is in line with previous reports conducted on different organisms showing that the frequencies of UAA and UGA STOP codons are strongly dependent on the GC content (Povolotskaya et al. 2012; Korkmaz et al. 2014; Belinky et al. 2018).

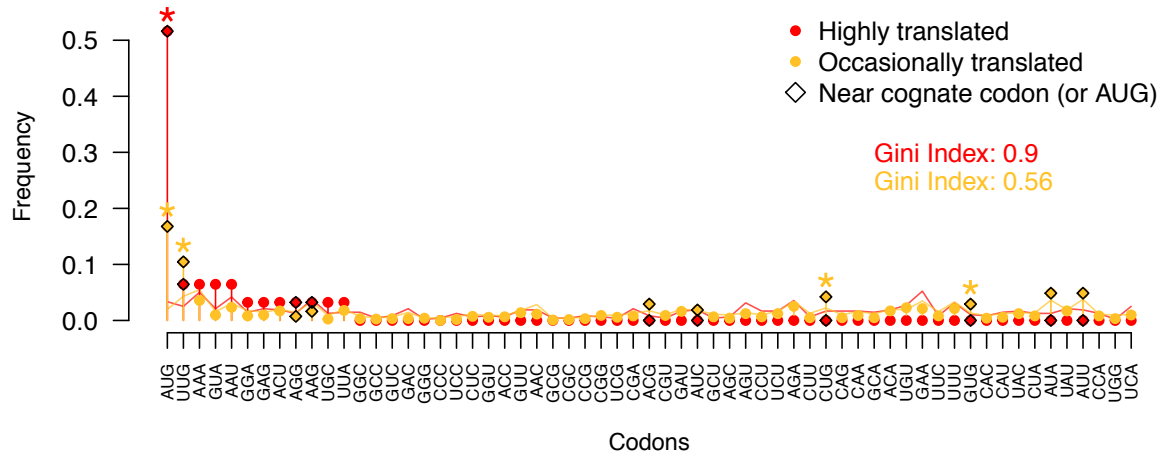


Figure 3.5. Frequencies of the 61 codons at the first translated position for highly translated IGORFs (red) and occasionally translated ones (yellow). Gini indexes which reflect the statistical dispersion of the 61 codons at the first translated position are given for highly and occasionally translated IGORFs in red and yellow respectively. Gini index values range from 0 to 1 and high values reflect the fact that the first translated positions are enriched in specific codons, particularly, AUG and other NUG ones. Codons that are significantly observed at the first translated position compared to the other translated positions are indicated with a star (one proportion z-test, $P < 5 \times 10^{-2}$). Near-cognate codons are indicated with diamonds.

Translated and ancestral IGORFs display intermediate properties between IGORFs and CDS

Figure 3.6A-D shows the boxplot distributions of the sizes of the sequences, clusters and linkers of all ORF categories along with their number of clusters per sequence. The HCA cluster size remains invariant for all categories except for de novo genes. In contrast, highly translated IGORFs, ancestral ones and de novo genes overall display for most properties, intermediate values between IGORFs and CDS.

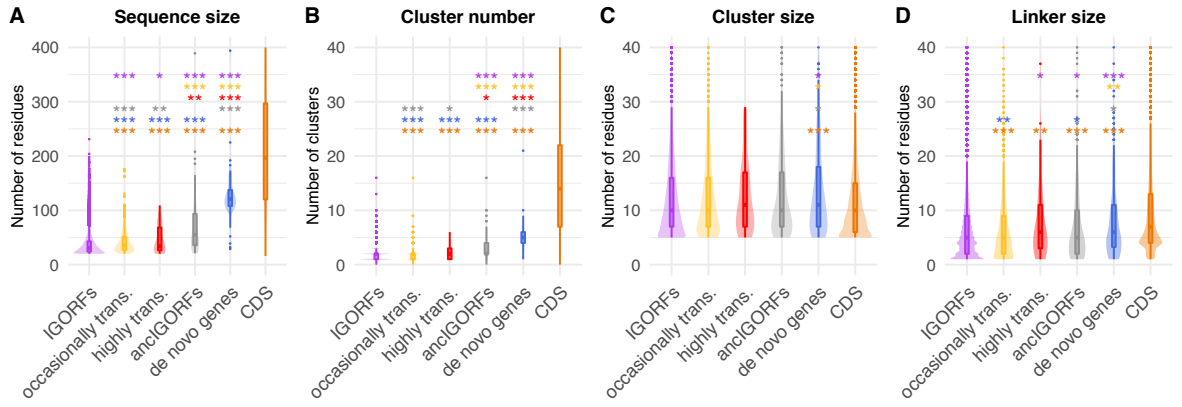


Figure 3.6. Continuum of sequence and structural properties between the different ORF categories. Comparison of **(A)** the sequence size, **(B)** cluster number, **(C)** cluster sizes, and **(D)** linker sizes for each ORF category (IGORFs in purple, occasionally translated IGORFs in yellow, highly translated IGORFs in red, ancIGORFs in grey, de novo genes in blue and CDS in orange). The p-values were computed with the Mann-Whitney *U* test (one-sided for (A), (B), (D), and two-sided for (C)). Asterisks denote level of significance: *p < 5×10^{-2}, **p < 1×10^{-2}, ***p < 1×10^{-3}. For each plot, the color of the asterisks indicates the ORF category used for the comparison. The exact p-values are given in Supplemental Tables S3.1-3.4.

In particular, the highly translated IGORFs, and the ancIGORFs are both longer than IGORFs (Mann-Whitney *U* test, P = 3.4×10^{-2} and 1.3×10^{-22} respectively), display slightly longer linkers (Mann-Whitney *U* test, P = 2.6×10^{-2} and 1.8×10^{-2}), and higher GC contents (41.9%, 38%, and 36.1% for ancIGORFs, highly translated IGORFs, and IGORFs respectively). In order to understand whether the increase in linker size could be explained by the increase in ORF length or GC content, we generated artificial IGORFs with nucleotide compositions of IGORFs and size distribution of ancIGORFs or highly translated IGORFs respectively. Artificial IGORFs with ancIGORF lengths exhibit linkers of similar size to those of IGORFs (Mann-Whitney *U* test, P = 2×10^{-1}) showing that the increase in linker sizes observed for ancIGORF cannot be explained by their larger size (Supplemental Fig. S3.11). However, the artificial linkers are shorter than those of ancIGORFs (Mann-Whitney *U* test, P =6×10^{-4}) suggesting that the effect can be attributed to the nucleotide composition of ancIGORFs. Indeed, scrambling the ancIGORF nucleotides results in linker sizes similar to those of ancIGORFs suggesting that the sole GC content of ancIGORFs is sufficient to generate long linkers. A similar trend is observed for highly translated IGORFs, though the effect is less pronounced (Supplemental Fig. S3.11). More generally, if for extreme hydrophobic and hydrophilic contents, the

sequence length has a substantial impact on cluster and linker sizes, for intermediate hydrophobic contents such as those of all ORF categories including CDS, the sequence length has no or small effect (Supplemental Fig. S3.12). As a matter of fact, artificial IGORFs with CDS sizes and IGORF nucleotide compositions are characterized by shorter linkers than those of real and scrambled CDS (Mann-Whitney *U* test, $P = 7.1 \times 10^{-8}$ and 2×10^{-4} respectively) (Supplemental Fig. S3.13). All these results reveal that the size of linkers results from a subtle combination of sequence length, GC content, and finally, of the resulting amino acid composition (Supplemental Fig. S3.12; Supplemental Fig. S3.13; Supplemental Fig. S3.14).

Discussion

In this work, we showed that the noncoding genome encodes the raw material for making proteins. In particular, we showed the widespread existence in the noncoding genome of the elementary building blocks of protein structures. Hydrophobic clusters in noncoding sequences display sizes similar to those observed in CDS. In contrast, CDS are enriched in longer linkers which probably contribute to optimize the local arrangements of secondary structures, provide flexibility to proteins, and specificity in protein interactions. This observation is in line with several studies reporting a central role to loops in protein function and structural innovation (Blouin et al. 2004; Tendulkar et al. 2004; Espadaler et al. 2006; Papaleo et al. 2016). Like Schmitz et al. (2018), we stipulate that the increase in intrinsic structural disorder observed for old genes in Carvunis et al. (2012), is related to the fact that CDS are characterized by longer linkers, thereby inducing an increase in the disorder score. As a matter of fact, most CDS display HCA scores similar to those of globular proteins, with low disorder propensities (Figure 3.2). Overall, we showed an enrichment in polar and charged residues for CDS which may be accompanied by an increase in specificity of protein folds and interactions through the optimization of the folding and assembly processes (Lumb and Kim 1995). De novo genes display a GC content similar to the one of CDS while their amino acid composition is rather IGORF-like. The effect is even stronger for ancIGORFs which are characterized by the highest GC content of all ORF categories while displaying an IGORF-like amino acid composition. This suggests an important role for the GC content in de novo gene

emergence, as reported in Vakirlis et al. (2018). We can hypothesize that the amino acid composition is optimized afterward while maintaining the GC content through the structure of the genetic code.

Nevertheless, how a noncoding sequence becomes coding remains unclear. In this work, we propose the IGORFs as potential elementary modules of protein birth and evolution. IGORFs could serve as starting points for de novo gene emergence or could be combined together, thus increasing protein sizes, contributing to protein modularity, and leading to more complex protein architectures. They resonate with the short protein fragments, reported so far, that result from different protein structure decompositions with the aim of partitioning protein structures into universal basic units of folding, folds and/or function (Berezovsky et al. 2000, 2001; Lamarine et al. 2001; Papandreou et al. 2004; Alva et al. 2015; Postic et al. 2017; Nepomnyachiy et al. 2017). The sizes of these structural fragments, overall, range from 25 to 35 residues with the exception of the "themes" (Kolodny et al. 2021) (average of 49 residues) and precisely recall those of IGORFs. Additionally, we showed that IGORFs encompass all the protein fold potential diversity observed in CDS. A majority of IGORFs encode peptides predicted as foldable while an important fraction displays high HCA scores and aggregation propensities. Some of the latter, though not the majority (28%), are predicted with at least one TM domain and may “safely” locate in membranes as proposed in Vakirlis et al. (2020a). The impact of the other high HCA score IGORFs on the cell deserves further investigation. Nevertheless, we can hypothesize that if produced, most of the time, their concentration will not be sufficient to be deleterious (Langenberg et al. 2020). Indeed, it seems that for CDS, a certain degree of aggregation is tolerated at low concentration (Supplemental Fig. S3.15). On the other hand, although IGORFs with intermediate HCA scores may exhibit a certain propensity for aggregation, we can hypothesize that these aggregation-prone regions, under the hydrophobic effect, may play a role in their capacity to fold, in line with the hypothesis of an amyloid origin of the globular proteins (Greenwald and Riek 2012; Langenberg et al. 2020). We hypothesize that the balanced equilibrium of hydrophobic and hydrophilic residues observed for these IGORFs (39.1% of hydrophobic residues to be compared with the 50.8% observed for high HCA score IGORFs) may render possible the burying of aggregation-prone regions and the exposure of hydrophilic residues that is accompanied by an increase in foldability. We can hypothesize that, if produced, these

IGORFs could form small compact structures and/or could be stabilized through oligomerization or interactions with other proteins. Precisely, we showed that ancestral IGORFs predating de novo gene emergence are not IGORF-like, but rather enriched in sequences with a high propensity for foldability. Nevertheless, we can reasonably hypothesize that de novo peptides struggle to fold into a well-defined and specific 3D structure as shown with the young de novo genes *BSC4* and *goddard* identified in the *S. cerevisiae* and *D. melanogaster* lineages respectively (Namy et al. 2003; Bungard et al. 2017; Lange et al. 2021). In particular, Bungard et al. (2017) reported that the Bsc4 protein folds partially to an ordered structure that is unlikely to be unfolded according to Circular Dichroism spectra and bioinformatic analyses. However, despite this "rudimentary" fold, they show through Mass Spectrometry and denaturation experiments that Bsc4 is able to form compact oligomers. Its hydrophobic residue content (38%) is higher than the one of CDS (33%) and typical of foldable IGORFs (39%). Whether this may be related to its "rudimentary" fold is questionable. We can hypothesize that the specificity of the Bsc4 structure will increase during evolution through amino acid substitutions toward hydrophilic residues.

Altogether, these results enable us to propose a model (Figure 3.7) which gives a central role to IGORFs in de novo gene emergence and to a lesser extent in protein evolution, thus completing the large palette of protein evolution mechanisms such as duplication events, horizontal gene transfer, domain shuffling... This model unifies two evolutionary processes that are usually addressed separately: the origin of novel genes and the elongation and thus evolution of pre-existing proteins, through IGORFs as elementary molecular modules widespread in noncoding regions. Once an IGORF is selected (Figure 3.7A), it can be subjected to different mutational events such as nucleotide substitutions or indels. In our model, multiple rounds of nucleotide substitutions are expected to change the amino acid landscape of the selected IGORF as shown with the enrichment of CDS in hydrophilic residues. We can hypothesize that mutations of hydrophobic residues towards hydrophilic ones can disrupt weak clusters into linkers or can switch cluster extremities into linker extremities, thereby increasing the size of linkers (Figure 3.7B). Besides, we hypothesize that the selected IGORF can elongate through indels and/or STOP codon mutations, thus incorporating a neighboring IGORF (Figure 3.7C). We hypothesize that the combination of two neighboring IGORFs through indels or STOP

codon mutations can lead to the creation of long linkers at the IGORFs' junction as observed in the example of the YMR153C-A de novo gene (Supplemental Fig. S3.16A). Similarly, the fusion of ancIGORFs can also give rise to long clusters as observed with the YPR126C de novo gene (Supplemental Fig. S3.16B), although it seems that long clusters have not been retained by selection as suggested by the CDS cluster size which is similar to the one of IGORFs. We showed with the reconstruction of 70 yeast de novo genes and in line with Zhang et al. (2019), that STOP codon mutations are less frequent than indels. Bartonek et al. (2020), reported that the hydrophobicity profiles of protein sequences remain invariant after frameshift events thanks to the interdependence of the three reading frames. Consequently, indels or frameshift events are most of the time expected to incorporate an IGORF that encodes a peptide with a hydrophobicity profile similar to that of the preexisting gene and may explain the fact that they are more frequent than STOP codon mutations. This suggests that the fold potential is a critical feature that needs to be conserved even in noncoding sequences, being preserved in +1, -1 phases through the structure of the genetic code. In addition, we showed that combining IGORFs encoding foldable peptides with IGORFs encoding disorder or aggregation-prone ones has low impact on the foldability of the resulting de novo proteins of the study. We can hypothesize that the newly integrated IGORFs will benefit from the structural properties of the preexisting IGORF network. Proteins can be seen as assemblies on an ancient protein core, whatever its evolutionary history, of either duplicated, shuffled domains or de novo translated products encoded by neighboring IGORFs (Figure 3.7D). Overall, in line with recent evolutionary fragment-based protein design developments, this model offers a rational framework for designing novel chimeric proteins by combining small elementary modules with specific structural properties (Höcker 2014; Berezovsky 2019; Yin et al. 2021; Ferruz et al. 2021; Bornberg-Bauer et al. 2021).

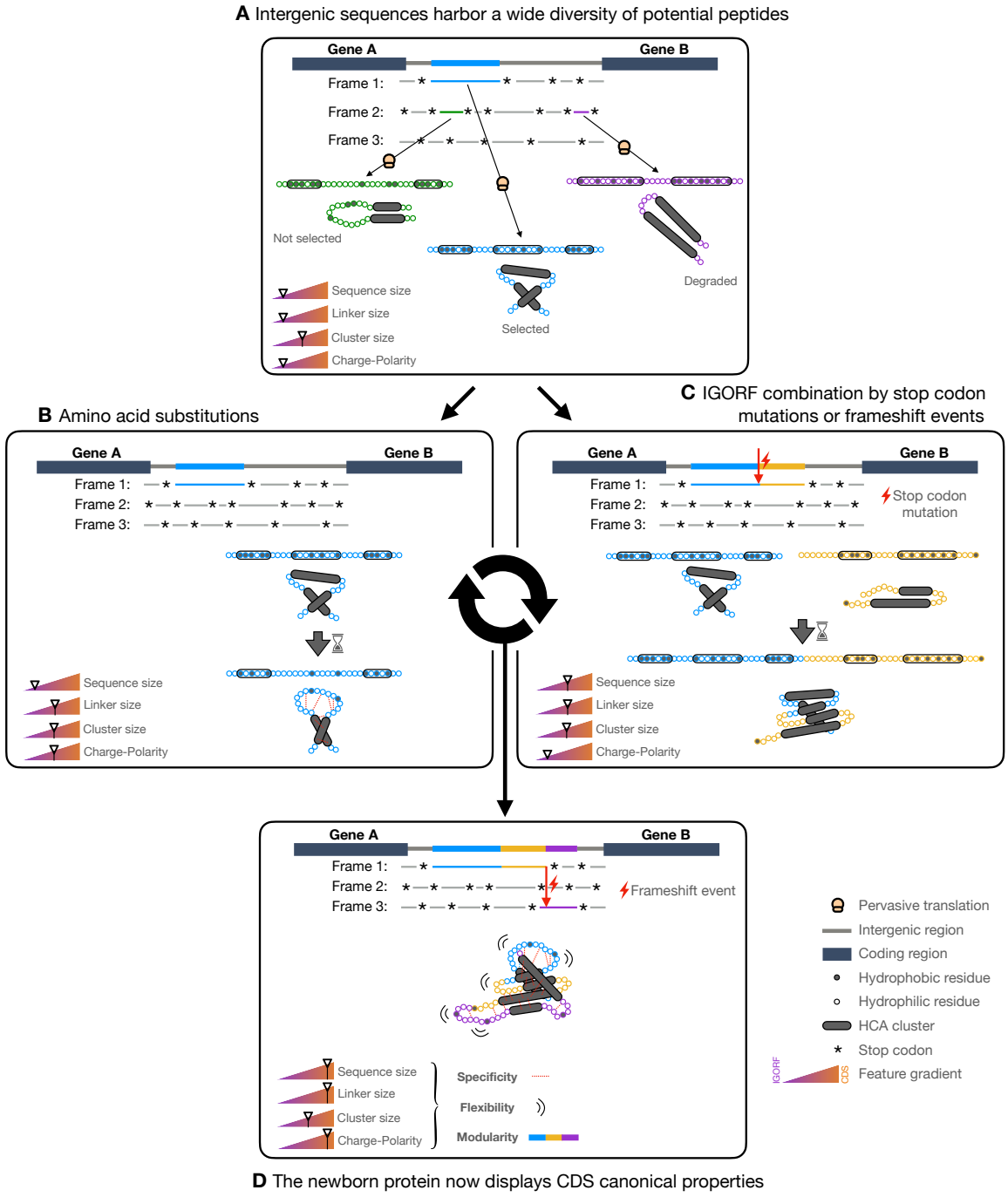


Figure 3.7. Model of de novo gene emergence and protein evolution with IGORFs as elementary structural modules. **(A)** IGORFs encode a wide diversity of peptides from disorder-prone to aggregation-prone ones, among which, a vast amount is expected to be able to fold in solution. Upon pervasive translation, some peptides that can be deleterious or not will be degraded right away. Among the others, the blue one will confer an advantage to the organism and will be further selected, thus providing a starting point for de novo gene birth. **(B)** The starting point IGORF, once selected, is subjected to amino acid substitutions thereby increasing the overall proportion of hydrophilic residues of the encoded peptide. In the present

case, this induces (i) the disruption of the second cluster resulting in the increase of the size of the central linker and (ii) the establishment of specific interactions between hydrophilic residue (red dots) which increase the specificity of the folding process and the resulting fold. **(C)** The STOP codon of the starting point IGORF can be mutated into an amino acid, thereby adding the yellow IGORF to the preexisting selected IGORF and elongating its size. **(D)** After multiple events of amino acid substitutions and IGORF combinations through STOP codon mutations or indels, we obtain a protein which displays the canonical features of CDS (i.e. long sequences, long linkers, enrichment in polar and charged residues) which enable the optimization of its flexibility, the increase in specificity of its folding process, 3D fold, and interactions and finally participate along with domain shuffling or duplication events in the modular architecture of genuine proteins. One should notice that although the figure focuses on de novo gene emergence, this model can also apply to already existing proteins.

Our model is supported by previous observations which show that (i) de novo genes are shorter than old ones (Wolf et al. 2009; Tautz and Domazet-Lošo 2011), (ii) the size of de novo gene exons are similar to those of old genes (Neme et al. 2017; Palmieri et al. 2014; Schlötterer 2015), and (iii) novel domains are generally observed in the C-terminal regions (Bornberg-Bauer et al. 2015; Klasberg et al. 2018). Nevertheless, a lot of questions regarding the mechanisms predating the selection of an IGORF remain open. Figure 3.6 displays a continuum in the presented properties between IGORFs and CDS that recalls the proto-gene model proposed by Carvunis et al. (2012), though the continuity between the translated IGORFs and the ancestral ones is to be demonstrated. Whether the high translation signal of highly translated IGORFs derives directly from the acquisition of a Methionine or whether it derives from previously occasionally translated IGORFs that have optimized their translational activity remains unclear. Similarly, the fate of highly translated IGORFs and their relationship with ancIGORFs are to be further characterized. Indeed, among the population of highly translated ORFs, some of them may give rise to future novel genes, thereby constituting today, the ancestral IGORFs of tomorrow, while others may be short-lived in evolutionary history. Finally, the increase in sequence and linker sizes observed between the different ORF categories opens several questions. We showed that the increase in linker size for ancIGORFs can be explained by their GC content and finally their amino acid composition. Precisely, ancIGORFs display a higher GC content than IGORFs (41.9% and 36.1% respectively) suggesting a role for GC-rich genomic regions in de novo gene properties and emergence as reported in previous studies (Basile et al. 2017; Vakirlis et al. 2018). Whether this

increase in GC content is accompanied by an increase in sequence length (STOP codons are AT-rich), linker size, and finally in foldability is a very interesting question that deserves further study. Indeed, it is still unknown whether the linker size is simply the consequence of the enrichment of CDS in hydrophilic residues and the increase in protein size or whether harboring long linkers is accompanied by an increase in foldability and is thus a selected criterion. Finally, all these results highlight an intimate relationship between sequence length, GC content and amino acid composition, whose combination is directly related to the size of linkers and clusters and finally to the foldability of the resulting product. Which one or which combination has driven the evolution of CDS? Our results cannot enable us to conclude. Nevertheless, the function of a protein derives directly from its structure and interactions, and can be, more generally, related to the concepts of stability, specificity and diversity. These concepts are in turn related to the equilibrium between hydrophobic and hydrophilic residues, protein modularity and finally, protein size which may altogether shape the linker and cluster size of proteins.

In this work, we propose a model that covers the genesis of all the diversity of the structural states observed in current proteins. If IGORFs encoding foldable peptides seem to be more likely to give rise to novel genes, disordered or aggregation-prone de novo proteins may emerge occasionally (Figure 3.4B). They are most of the time (79%) associated with ancIGORFs expected to encode disordered or aggregation-prone peptides as well, suggesting that the structural properties of de novo proteins are already encoded in the ancestral peptide they originate from. Whether the fold potential of a starting point IGORF conditions the structural properties of the resulting de novo protein is an exciting question that deserves further study. Indeed, we can hypothesize that once selected, an IGORF can elongate over time through the incorporation of neighboring IGORFs, provided that the latter do not affect the fold potential of the preexisting protein. In accordance with Vakirlis et al. (2020a), we can reason that once a starting point IGORF is selected, it engenders novel selected effects which in turn, increase the constraints exerted on it and subsequently reduce the possibility of future changes. It is thus tempting to speculate that the structural properties of the peptide encoded by the starting point IGORF will be retained during evolution through the elimination of the deleterious IGORFs' combinations. All these observations suggest that the diversity of the structural states observed in current proteins has been originally inherited from the

diversity of the fold potential already encoded in the noncoding genome. If and how the noncoding genome can account for the structural diversity of proteins is another exiting question that deserves another study.

Methods

Datasets:

CDS and IGORFs: The CDS were extracted from the genome of *Saccharomyces cerevisiae* S288C according to the genome annotation of the S*accharomyces* Genome database (Cherry et al. 2012). All unannotated ORFs of at least 60 nucleotides, no matter if they start with an AUG codon, were extracted from the 16 yeast chromosomes. We only retained ORFs that are free from overlap with another gene or that partially overlap with a gene if the non-overlapping region is more than 70% of the IGORFs sequence. ***Datasets of reference:*** The disorder dataset consists of 731 disordered regions extracted from intrinsically disordered proteins of the DisProt database (Hatos et al. 2020), that were used for the calibration of HCAtk (Bitard-Feildel and Callebaut 2018a). The globular dataset consists of 559 globular proteins extracted from the Protein Data Bank (Berman et al. 2000; Burley et al. 2021) that were used for the calibration of IUPred (Dosztanyi et al. 2005; Mészáros et al. 2009; Dosztányi 2018; Mészáros et al. 2018). The transmembrane regions dataset gathers 1269 transmembrane regions extracted from the transmembrane proteins contained in the PDBTM database (Tusnády et al. 2004, 2005; Kozma et al. 2012). We only retained transmembrane segments longer than 20 amino acids corresponding to the minimum size of an IGORF. ***Random noncoding genome:*** Intergenic regions were concatenated, and their nucleotides were scrambled. Then random IGORFs of at least 60 nucleotides were extracted as explained above. ***Scrambled sequences:*** scrambled sequences were generated by shuffling the nucleotides of the ORFs of interest. When an in-frame STOP codon was generated, its 3 nucleotides were randomized until they did not lead to a STOP codon. ***Artificial IGORFs:*** we generated artificial sequences of fixed size (e.g. size of CDS) by drawing nucleotides according to the nucleotide composition of IGORFs.

Estimation of the fold potential, the aggregation, disorder and TM propensities

The foldability potential was estimated using a score derived from the HCA (Hydrophobic Cluster Analysis) approach using the HCAtk program (Bitard-Feildel and Callebaut 2018a; Bitard-Feildel et al. 2018), while the disorder and aggregation propensities were assessed with IUPred and TANGO respectively (Supplemental Methods) (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a, Dosztanyi et al. 2005; Mészáros et al. 2009; Dosztányi 2018; Mészáros et al. 2018). The presence of TM domains was predicted with TMHMM (Krogh et al. 2001).

Protein abundances and amino acid propensities

Protein abundance data were extracted from the PaxDB database (Wang et al. 2012). In order to depict the impact of the avoidance of nonspecific interactions with the ribosome, we only retained cytoplasmic proteins as annotated in UniProt (The UniProt Consortium 2019). The propensity of an amino acid *i* to be found in a CDS cluster is defined by the log ratio of the frequencies of the amino acid *i* in CDS clusters versus IGORF clusters as follows:

$$\text{propensity}(\text{aa}_\text{i} \text{ in CDS clusters}) = \log_{10}\left(\frac{\text{freq}(\text{aa}_\text{i})_{\text{ in CDS clusters}}}{\text{freq}(\text{aa}_\text{i})_{\text{ in IGORF clusters}}}\right)$$

Reconstruction of Ancestral IGORFs

To reconstruct the ancIGORFs of *S. cerevisiae*, we used the genomes of the neighboring species *S. paradoxus* (Durand et al. 2019), *S. arboricola* (Yue et al. 2017), *S. mikatae, S. kudriavzevii,* and *S. uvarum* (Scannell et al. 2011). Based on four independent studies which each listed de novo genes of the *S. cerevisiae* genome, we retained all de novo genes identified in at least two studies (Carvunis et al. 2012; Lu et al. 2017; Vakirlis et al. 2018; Wu and Knudson 2018). This led to a total of 171 de novo genes among which we retained those for which we were able to identify at least two additional homologous sequences in the neighboring species among which, at least one had to be noncoding in order to reconstruct the corresponding nongenic region in the ancestor (Supplemental Table S6). Therefore, we searched for the orthologous genes of the 70 de novo genes in the

neighboring species using BLASTP (evalue < 1×10^{-2}) (Supplemental Fig. S3.7A). Then, based on the species tree and starting from the branch of *S. cerevisiae*, we traced back to the root and identified the first node branching with a branch for which no orthologous gene had been detected (yellow circle in Supplemental Fig. S3.7A). We hypothesize that the corresponding locus in the ancestor was still nongenic. We searched for the corresponding nongenic regions in the remaining species with TBLASTN (evalue < 1×10^{-2}). Following the protocol described by Vakirlis and McLysaght (2019), the resulting homologous nucleotide sequences and orthologous de novo genes were subsequently aligned with MACSE v2.05 (Ranwez et al. 2011, 2018) and the corresponding phylogenetic tree was constructed with PhyML (Guindon et al. 2010). The multiple sequence alignment and its corresponding tree were given as input to PRANK (Löytynoja and Goldman 2010) for the reconstruction of the corresponding ancestral nongenic nucleotide sequence (Supplemental Fig. S3.7B,C). Finally, the ancestral nucleotide sequences were translated into the three reading frames. The resulting IGORFs were then aligned with the de novo gene of *S. cerevisiae* with LALIGN (Huang and Miller 1991) those sharing a homology with it were retained (Supplemental Fig. S3.7D).

Ribosome Profiling analyses

Ribosome profiling datasets: we used five ribosome profiling datasets of wild type *S. cerevisiae*, two of which were generated in the present study (Supplemental Methods) (GEO accession number GSE173861, samples GSM5282046 and GSM5282047). The three others were taken from Radhakrishnan et al. (2016) (GEO accession number GSE81269, samples GSM2147982 and GSM2147983) and Thiaville et al. (2016) (GEO accession number GSE72030, sample GSM1850252). ***Selection of RPF (Ribosome Protected Fragments):*** Ribosome profiling reads were mapped on the genome of *S. cerevisiae* S288C using Bowtie (Langmead et al. 2009). For this study, we only kept the 28-mers since on average 90% of them were mapped on a CDS in the correct reading frame (Supplemental Fig. S3.17). ***Periodicity:*** The periodicity is calculated using a metagene profile. It provides the number of footprints relative to all annotated start codons in a selected window. The metagene profile is obtained by pooling together all the annotated CDS and counting the number of RPFs at each nucleotide position. Supplemental Fig. S3.17 shows a clear accumulation of signal over the CDS, and a nice periodicity over the 100 first nucleotides. ***Identification of the occasionally translated***

IGORFs: we retained the IGORFs with at least 10 reads in at least one dataset. ***Identification of the highly translated IGORFs:*** we kept the IGORFs with at least 30 reads in at least two datasets for which the fraction of in-frame reads was higher than 0.8.

Statistical analyses

All statistical analyses that aimed at comparing distributions were performed in R (4.0.3) (Team R Core 2020) using the Kolmogorov–Smirnov test (two-sided) when comparing whether the HCA score distributions are statistically different and the Mann Whitney *U* test for the comparison of the median cluster size, linker size, sequence size and cluster number distributions (bilateral test for the comparison of cluster sizes and unilateral test for the other properties). We used the one proportion z-test for the comparison of the proportion of disordered, foldable or aggregation prone sequences between different ORF categories. In order to circumvent the *p-value problem* inherent to large samples (Lin et al. 2013), tests were performed iteratively 1000 times on samples of 500 individuals randomly chosen from the initial sample when it was larger than 500 individuals. The averaged p-value over the 1000 iterations was subsequently calculated.

Data Access

The raw ribosome profiling data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE173861. Raw and calculated data along with codes to reproduce analyses and figures are available as Supplemental Code 1, and the programs to extract the IGORFs and estimate their structural properties (ORFtrack and ORFold) are available in the ORFMine package as Supplemental Code 2 and on GitHub (https://github.com/i2bc/ORFmine).

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

CP work was supported by a French government fellowship.

Author contributions: CP, MR, IH performed research, CP, MR, IH, ON, AL analyzed data. CP, AL designed research. CP, IC, JCG, ON, OL, AL wrote the paper. AL conceived the project

References

Alva V, Söding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**: e09410.

Bartonek L, Braun D, Zagrovic B. 2020. Frameshifting preserves key physicochemical properties of proteins. *Proc Natl Acad Sci* **117**: 5907–5912.

Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. *PLoS Comput Biol* **13**: e1005375.

Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Sci Rep* **8**: 1–11.

Berezovsky IN. 2019. Towards descriptor of elementary functions for protein design. *Curr Opin Struct Biol* **58**: 159–165.

Berezovsky IN, Grosberg AY, Trifonov EN. 2000. Closed loops of nearly standard size: common basic element of protein structure. *Febs Lett* **466**: 283–286.

Berezovsky IN, Kirzhner VM, Kirzhner A, Trifonov EN. 2001. Protein folding: looping from hydrophobic nuclei. *Proteins Struct Funct Bioinforma* **45**: 346–350.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res* **28**: 235–242.

Bitard-Feildel T, Callebaut I. 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep* **7**: 1–13.

Bitard-Feildel T, Callebaut I. 2018. HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences. *bioRxiv* 249995.

Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. 2015. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis." *Biochimie* **119**: 244–253.

Bitard-Feildel T, Lamiable A, Mornon J, Callebaut I. 2018. Order in disorder as observed by the "hydrophobic cluster analysis" of protein sequences. *Proteomics* **18**: 1800054.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat Commun* **12**: 1–13.

Blouin C, Butt D, Roger AJ. 2004. Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. *Protein Sci* **13**: 608–616.

Bornberg-Bauer E, Alba MM. 2013. Dynamics and adaptive benefits of modular protein evolution. *Curr Opin Struct Biol* **23**: 459–466.

Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol* **68**: 175–183.

Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from 'dark genomic matter'by 'grow slow and moult.' *Biochem Soc Trans* **43**: 867–873.

Bresler SE, Talmud D. 1944. On the nature of globular proteins. *CR Acad Sci USSR* **43**: 310–314.

Brooks DJ, Fresco JR. 2003. Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. *Gene* **303**: 177–185.

Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MH. 2017. Foldability of a natural de novo evolved protein. *Structure* **25**: 1687–1696.

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, et al. 2021. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **49**: D437–D451.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370–374.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–D705.

Cuevas MVR, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, Lanoix J, Côté C, Staudt LM, Lemieux S, et al. 2021. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* **34**: 108815.

Dosztányi Z. 2018. Prediction of protein disorder based on IUPred. *Protein Sci* **27**: 331–340.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**: 827–839.

Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res* **29**: 932–943.

Ekman D, Elofsson A. 2010. Identifying and quantifying orphan protein sequences in fungi. *J Mol Biol* **396**: 396–405.

Espadaler J, Querol E, Aviles FX, Oliva B. 2006. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics* **22**: 2237–2243.

Faure G, Callebaut I. 2013a. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput Biol* **9**.

Faure G, Callebaut I. 2013b. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics* **29**: 1726–1733.

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**: 1302–1306.

Ferruz N, Noske J, Höcker B. 2021. Protlego: a Python package for the analysis and design of chimeric proteins. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab253 (Accessed August 4, 2021).

Foy SG, Wilson BA, Bertram J, Cordes MH, Masel J. 2019. A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. *Genetics* **211**: 1345–1355.

Ganesan A, Siekierska A, Beerten J, Brams M, Van Durme J, De Baets G, Van der Kant R, Gallardo R, Ramakers M, Langenberg T, et al. 2016. Structural hot spots for the solubility of globular proteins. *Nat Commun* **7**: 1–15.

Greenwald J, Riek R. 2012. On the possible amyloid origin of protein folds. *J Mol Biol* **421**: 417–426.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, Bassot C, Benítez GI, Bevilacqua M, Chasapi A, et al. 2020. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res* **48**: D269–D276.

Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. *J Mol Evol* **88**: 382–398.

Höcker B. 2014. Design of proteins from smaller fragments—learning from evolution. *Curr Opin Struct Biol* **27**: 56–62.

Huang X, Miller W. 1991. A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* **12**: 337–357.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789–802.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. 2018. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J* **285**: 2605–2625.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752–1759.

Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. 2021. Bridging themes: short protein segments found in different architectures. *Mol Biol Evol* **38**: 2191–2208.

Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem* **289**: 30334–30342.

Kozma D, Simon I, Tusnady GE. 2012. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res* **41**: D524–D529.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Lamarine M, Mornon J-P, Berezovsky IN, Chomilier J. 2001. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell Mol Life Sci CMLS* **58**: 492–498.

Lamiable A, Bitard-Feildel T, Rebehmed J, Quintus F, Schoentgen F, Mornon J-P, Callebaut I. 2019. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* **167**: 68–80.

Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. 2021. Structural and functional characterization of a putative de novo gene in Drosophila. *Nat Commun* **12**: 1–13.

Langenberg T, Gallardo R, van der Kant R, Louros N, Michiels E, Duran-Romaña R, Houben B, Cassio R, Wilkinson H, Garcia T, et al. 2020. Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins. *Cell Rep* **31**: 107512.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: 1–10.

Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci* **109**: 20461–20466.

Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the origin of de novo genes in Arabidopsis thaliana populations. *Genome Biol Evol* **8**: 2190–2202.

Lin M, Lucas Jr HC, Shmueli G. 2013. Research commentary—too big to fail: large samples and the p-value problem. *Inf Syst Res* **24**: 906–917.

Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* **342**: 345–353.

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**: 579.

Lu T-C, Leu J-Y, Lin W-C. 2017. A comprehensive analysis of transcript-supported de novo genes in Saccharomyces sensu stricto yeasts. *Mol Biol Evol* **34**: 2823–2838.

Lumb KJ, Kim PS. 1995. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**: 8642–8648.

Macossay-Castillo M, Marvelli G, Guharoy M, Jain A, Kihara D, Tompa P, Wodak SJ. 2019. The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity. *J Mol Biol* **431**: 1650–1670.

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**: W329–W337.

Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* **5**: e1000376.

Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. *PloS One* **7**: e48650.

Namy O, Duchateau-Nguyen G, Hatin I, Hermann-Le Denmat S, Termier M, Rousset J. 2003. Identification of stop codon readthrough genes in Saccharomyces cerevisiae. *Nucleic Acids Res* **31**: 2289–2296.

Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat Ecol Evol* **1**: 1–7.

Nepomnyachiy S, Ben-Tal N, Kolodny R. 2017. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc Natl Acad Sci* **114**: 11703–11708.

Nielly-Thibault L, Landry CR. 2019. Differences between the raw material and the products of de novo gene birth can result from mutational biases. *Genetics* **212**: 1353–1366.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. *elife* **3**: e01311.

Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. 2016. The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev* **116**: 6391–6423.

Papandreou N, Berezovsky IN, Lopes A, Eliopoulos E, Chomilier J. 2004. Universal positions in globular proteins: From observation to simulation. *Eur J Biochem* **271**: 4762–4768.

Postic G, Ghouzam Y, Chebrek R, Gelly J-C. 2017. An ambiguity principle for assigning protein structural domains. *Sci Adv* **3**: e1600552.

Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biol Direct* **7**: 1–13.

Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* **167**: 122–132.

Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol* **35**: 2582–2584.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS One* **6**: e22594.

Rousseau F, Schymkowitz J, Serrano L. 2006a. Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struct Biol* **16**: 118–126.

Rousseau F, Serrano L, Schymkowitz JW. 2006b. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* **355**: 1037–1047.

Scannell D, Zill O, Rokas A, Payen C, Dunham M, Eisen M, Rine J, Johnston M, Hittinger C. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda). 2011; 1 (1): 11–25. *Genet Soc Am*.

Schavemaker PE, Śmigiel WM, Poolman B. 2017. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. *Elife* **6**: e30084.

Schlötterer C. 2015. Genes from scratch–the evolutionary fate of de novo genes. *Trends Genet* **31**: 215–219.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol* **2**: 1626–1632.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.

Team R Core RC. 2020. R: A language and environment for statistical computing. https://www.R-project.org/.

Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP. 2004. Clustering of protein structural fragments reveals modular building block approach of nature. *J Mol Biol* **338**: 611–629.

Thiaville PC, Legendre R, Rojas-Benítez D, Baudin-Baillieu A, Hatin I, Chalancon G, Glavic A, Namy O, de Crécy-Lagard V. 2016. Global translational impacts of the loss of the tRNA modification t6A in yeast. *Microb Cell* **3**: 29.

Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J, et al. 2017. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep* **7**: 1–9.

Trifonov E. 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *J Mol Biol* **194**: 643–652.

Tusnády GE, Dosztányi Z, Simon I. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res* **33**: D275–D278.

Tusnády GE, Dosztányi Z, Simon I. 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **20**: 2964–2972.

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506–D515.

Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines CP, Iannotta J, et al. 2020a. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun* **11**: 1–18.

Vakirlis N, Carvunis A-R, McLysaght A. 2020b. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Mol Biol Evol* **35**: 631–645.

Vakirlis N, McLysaght A. 2019. Computational prediction of de novo emerged protein-coding genes. In *Computational Methods in Protein Evolution*, pp. 63–81, Springer.

Van Oss SB, Carvunis A-R. 2019. De novo gene birth. *PLoS Genet* **15**.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* **11**: 492–500.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol* **1**: 1–6.

Wissler L, Godmann L, Bornberg-Bauer E. 2012. Evolutionary dynamics of simple sequence repeats across long evolutionary time scale in genus Drosophila. *Trends Evol Biol* **4**: e7–e7.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci* **106**: 7273–7280.

Wu B, Knudson A. 2018. Tracing the de novo origin of protein-coding genes in yeast. *MBio* **9**.

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7**: e1002379.

Yin M, Goncearenco A, Berezovsky IN. 2021. Deriving and using descriptors of elementary functions in rational protein design. *Front Bioinforma* **1**: 8.

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**: 913–924.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol* **3**: 679–690.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**: 769–772.

4 Study of sequence and structural properties of proteins along evolution

4.1 Introduction

In section 3, we studied and compared the sequence and structural properties between established coding ORFs (CDS) and noncoding intergenic ORFs (IGORFs) of the yeast. We showed that the IGORFs can encompass a large diversity of foldability potential while the CDS have evolved towards a more limited range of foldability which falls mostly in the range of globular and well folded proteins. We also showed that CDS are significantly longer than the IGORFs, hosting longer disordered regions (HCA linkers) while they both (CDS and IGORFs) present similar sizes of HCA clusters. Although we addressed the CDS as a single and homogeneous group of sequences, one should notice that the proteins of *S. cerevisiae* have appeared at different points of the evolutionary time and as a result, the proteins observed today have not been subjected to selection for the same length of time. Notably, older proteins have been subjected to selection for longer, compared to younger ones. If and how the evolutionary time has affected the sequence and structural properties of the CDS are interesting questions that we address in this section. CDS sequences code for functional proteins and most of them are found conserved among multiple organisms. Using homology detection methods, it becomes possible to assign the presence or absence of a CDS sequence from the genome of an organism and detect its last occurrence among a set of different organisms (Van Oss and Carvunis 2019). This procedure permits to assign a relative date to the CDS sequence which corresponds to the evolutionary timepoint of its last occurrence. This method is called genomic phylostratigraphy and has already been presented at the introduction of this manuscript as a method for orphan genes detection. Notably, ancient proteins will be present in all (or the majority) of the studied species while young proteins will be present only in a subset of species.

In this section, we aim at studying the evolution of several sequence and structural properties of the CDS along with the evolutionary time. To do so, the proteins of *S.*

cerevisiae were divided in ten distinct phylostrata (age groups) based on an already published phylostragraphy of the yeast (Wilson et al. 2017). The motivation of this analysis is the comparison of these properties among the different protein age groups and particularly between the youngest and the oldest phylostrata in order to capture any interesting tendencies established with evolution. Together with the already described sequence and structural predicted properties, we studied the evolution of additional properties of the yeast proteins such as the protein abundance in the cell, number of protein-protein interaction partners, their predicted structural domains and the secondary structure content of their 3D structures.

4.2 Methods

4.2.1 Age groups of *S. cerevisiae* proteins

For our analysis we used the relative ages of *S. cerevisiae* proteins as estimated with phylostratigraphy in the study of Wilson et al. (2017). Notably, genes taken in June 2014 from the Saccharomyces Genome Database (SGD) were subjected to a BLASTp search with an E-value threshold of 0.001 against the National Center for Biotechnology Information (NCBI) non-redundant protein sequences (nr) database. The most phylogenetically distant hit was used to place the gene into one of 10 defined phylostrata. The youngest phylostratum contained 562 proteins encoded by TRGs of the Saccharomyces genus, 304 of which were *S. cerevisiae* orphans. In addition, we created an eleventh group of genes corresponding to genes termed as “dubious” by the SGD and which are deemed unlikely to be real (Skrzypek and Hirschman 2011). In table 4.1 is presented the reparation of the *S. cerevisiae* proteins in the different age groups from the older to the younger phylostrata.

Table 4.1. Count of *S. cerevisiae* proteins per age group.

	Total number of proteins
Cellular Organisms	2575
Eukaryota	1578
Opisthokonta	127
Fungi	289
Dikarya	75
Ascomycota	140
Saccharomyceta	71
Saccharomycetales	366
Saccharomycetaceae	346
Saccharomyces	562
Dubious	545
Total	6649

4.2.2 Structural properties calculation

For the calculation of the foldability potential as well as the aggregation and disorder propensities we used the tools pyHCA (Faure and Callebaut 2013a, 2013b; Bitard-Feildel and Callebaut 2018a; Lamiable et al. 2019), Tango (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a) and IuPred (Mészáros et al. 2018; Dosztányi 2018; Erdős and Dosztányi 2020), respectively, through the ORFold tool which is part of the ORFmine package. See section 2 for details.

4.2.3 Proteins' abundance and Protein-Protein Interaction (PPI) partners

Protein abundance data were extracted from the PaxDB database (Wang et al. 2012) while the PPI of every protein were extracted from the BioGRID (version 4.4.200) database (Stark et al. 2006).

4.2.4 SCOP superfamilies annotation of the *S. cerevisiae* proteome

The SCOP superfamilies annotation of the *S. cerevisiae* proteome was downloaded from the SUPERFAMILY database (https://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/) (Gough et al. 2001). All the annotations had E-value < 0.01 and every SCOP superfamily was assigned to its corresponding SCOP class. For this analysis we focused on the five principal SCOP classes (all-α, all-β, α/β, α+β, multi-domain) and eliminated domains corresponding to other SCOP classes (Hubbard et al. 1997; Andreeva et al. 2014).

4.2.5 Dataset of *S. cerevisiae* protein 3D structures from the PDB

For the *S. cerevisiae* PDB protein structures dataset, we were based on the UniProt (UniProt Consortium 2019) annotation of the *S. cerevisiae* S288C strain. For every gene entry with PDB structures, we retained only one representative structure with the following preference: the X-ray against NMR structures and the X-ray with lowest resolution among all the X-ray cases. Like this we extracted 1346 PDB structures each one corresponding to a single *S. cerevisiae* gene. For more details about the repartition of the structures in the 10 age groups see table 4.3. For the secondary structures assignation on the PDB structures, the tool Stride (Heinig and Frishman 2004) was used.

4.2.6 Dataset of *S. cerevisiae* protein 3D structures models predicted by AlphaFold²

The 3D model structures of the proteome of *S. cerevisiae* predicted by AlphaFold² were downloaded from the Alphfold Protein Structure Database (https://alphafold.ebi.ac.uk) (Tunyasuvunakool et al. 2021). For more details about the repartition of the structures in the 10 age groups see table 4.3. For the secondary structures assignation on the AlphaFold² models, the tool Stride (Heinig and Frishman 2004) was used.

4.3 Results

Based on the phylostratigraphy presented in Wilson et al. (2017), the proteins of *S. cerevisae* were assigned with a relative age. Older proteins present homologs with more distantly related species while younger proteins are mostly restricted to close related

organisms. Out of the ten phylostrata, the dubious genes constituted a supplementary group which is not assigned with any age. Almost 83% (58/70) of the reconstructed de novo genes presented in section 3 correspond to this dubious category.

4.3.1 Evolution of the fold potential

In figure 4.1A is presented the HCA score distribution for the different yeast gene ages together with the ones of dubious genes and IGORFs. It is interesting to observe that the HCA score distribution becomes less wide and more centered with the evolutionary time. In order to verify that, we generated samples of 100 randomly selected proteins for every phylostratum and performed pairwise comparisons of the HCA scores variance for all the phylostrata with the F-test. This procedure was repeated 1000 times and an average p-value for the 1000 comparisons was calculated. The HCA scores of the oldest age group presented significantly different variance than any other age group (Two-sided F-test, all p-values $< 5 \times 10^{-2}$) reflecting that older proteins present more homogenous foldability potentials while younger proteins present a larger spectrum of foldabilities. Indeed, the HCA score distribution of the oldest-genes group (Cellular Organisms) is systematically different from the distribution of any other younger age group (Two-sided Kolmogorov Smirnov test, all p-values $< 2.2 \times 10^{-16}$). Notably, the HCA scores of all the protein age groups fall principally inside the foldable boundaries showing that foldability is an important trait that is constantly optimized during evolution. However, both the dubious and the young Saccharomyces TRGs present an intermediate behavior between IGORFs and older genes by acquiring a large range of foldability potentials. Interestingly, their foldability score distribution is not similar neither to IGORFs (Two-sided Kolmogorov Smirnov test, $P = 4 \times 10^{-3}$ and $P = 5 \times 10^{-3}$, respectively) nor to any other older age group (Two-sided Kolmogorov Smirnov test, all p-values $< 2.2 \times 10^{-16}$). Notably, they are both enriched in foldable sequences (80% and 79.7% for dubious and Saccharomyces TRGs, respectively) in comparison to IGORFs (66.6%) (one proportion z-test, $P = 4 \times 10^{-10}$ and $P = 3 \times 10^{-4}$ for dubious and Saccharomyces TRGs, respectively) such as the *S. cerevisiae* de novo genes presented in section 3. Figure 4.1B-C present the distribution of the disorder and the aggregation propensity, respectively, for the different yeast gene ages together with the ones of dubious genes and the IGORFs. One should notice that the distribution of these properties for the dubious and the Saccharomyces TRGs is significantly different from any other protein age group (Mann-Whitney U-test, all p-

values < 2.2×10^{-16}) while, on the contrary, they are statistically similar to the one of IGORFs (Mann-Whitney *U*-test, $P = 1 \times 10^{-1}$ and $P = 4 \times 10^{-1}$ for the disorder propensity; $P = 1 \times 10^{-1}$ and $P = 4 \times 10^{-1}$ for the aggregation propensity, for dubious and Saccharomyces TRGs, respectively). Pairwise comparisons between consecutive age groups for these two properties, revealed an overall stable tendency along with the evolutionary time with exception the older age groups which seem to present lower disorder propensity.

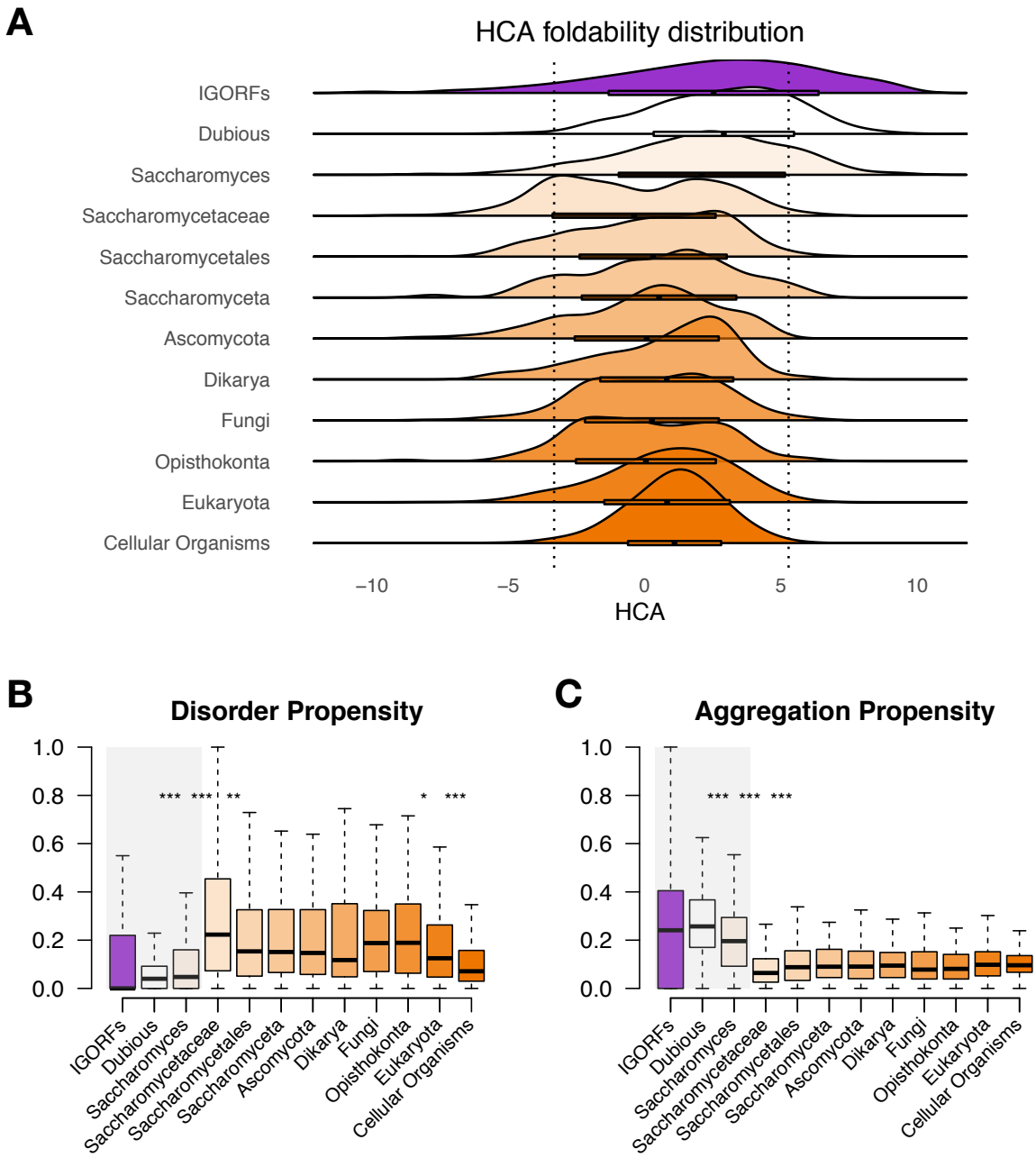


Figure 4.1. (A) Distribution of the HCA foldability score for the proteins of the ten *S. cerevisiae* age groups (presented in different shades of orange) as well as the IGORFs (presented in purple) and the dubious genes (presented in white). Dotted black lines delineate the boundaries of the low, intermediate and high HCA score categories reflecting the three categories of fold potential (i.e., disorder prone, foldable, or aggregation prone in solution). Horizontal bars correspond to the mean ± standard deviation of the distribution. **(B)** Boxplot distribution of the intrinsic disorder propensity of each protein age group (presented in different shades of orange) as well as the IGORFs (presented in purple) and the dubious genes (presented in white). **(C)** Boxplot distribution of the aggregation propensity of each protein age group (presented in different shades of orange) as well as the IGORFs (presented in purple) and the dubious genes (presented in white). Asterisks denote level of significance for the Mann-Whitney *U*-test for every consecutive pair: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$. The outliers of the boxplots are omitted for clarity.

Based on these results, the young yeast genes (Saccharomyces TRGs) present foldability score distribution which is different from both IGORFs and the rest of the older age groups, while their disorder and aggregation propensities are clearly different from the rest of the phylostrata and similar to the ones of the IGORFs. These support that the young genes might represent an intermediate state between noncoding ORFs and older genes. Notably, the same observations stand for the group of dubious genes making us wonder whether dubious genes in fact correspond to young genes.

4.3.2 Evolution of HCA clusters and linkers

In section 3, we showed that the sequence length, the HCA clusters' occurrences and the linkers' size of successive stages preceding the birth of de novo genes, display intermediate values between IGORFs and CDS supporting that the evolution of these properties somehow accompanies the emergence of novel genes. Precisely, long linkers were found to be specific to CDS while, on the contrary, the size of HCA clusters was invariant among the ORF categories (except the one of the de novo genes which are enriched in hydrophobic residues). This prompts us to investigate whether the increase in linker size, observed for CDS, was a continuous procedure over evolution or is a property that is fixed early in evolution.

We studied the sequence size and the HCA clusters number of the different gene groups, and we noticed the well-known tendency of proteins to elongate and acquire more HCA clusters with the evolutionary time (Figure 4.2A-B). Our finding that HCA clusters' sizes

do not differ significantly between IGORFs and CDS, still stands for all the yeast phylostrata and the dubious genes (Figure 4.2C), enforcing even more our initial results and supporting the idea of HCA clusters as elementary building blocks of proteins (Mann-Whitney *U*-test, all pairwise p-values > 0.05). On the other hand, the linkers' size between IGORFs and the proteins of any other age group (no matter younger, older or dubious) is significantly different (Figure 4.2D), with the IGORFs presenting smaller linker sizes, as already described in section 3. Interestingly, the group of young genes (Saccharomyces) presents intermediate linker sizes between the IGORFs and the Saccharomycetaceae age group (Mann-Whitney *U*-test, $P = 6 \times 10^{-8}$ and 3×10^{-2}, respectively) while the linker sizes among all the other age groups seems to be invariant along the evolutionary time. In addition, the dubious genes present intermediate linker sizes between IGORFs and the young yeast genes. Whether the dubious genes reflect an intermediate state between noncoding sequences and young yeast genes or just a younger subset of the young genes is an interesting question.

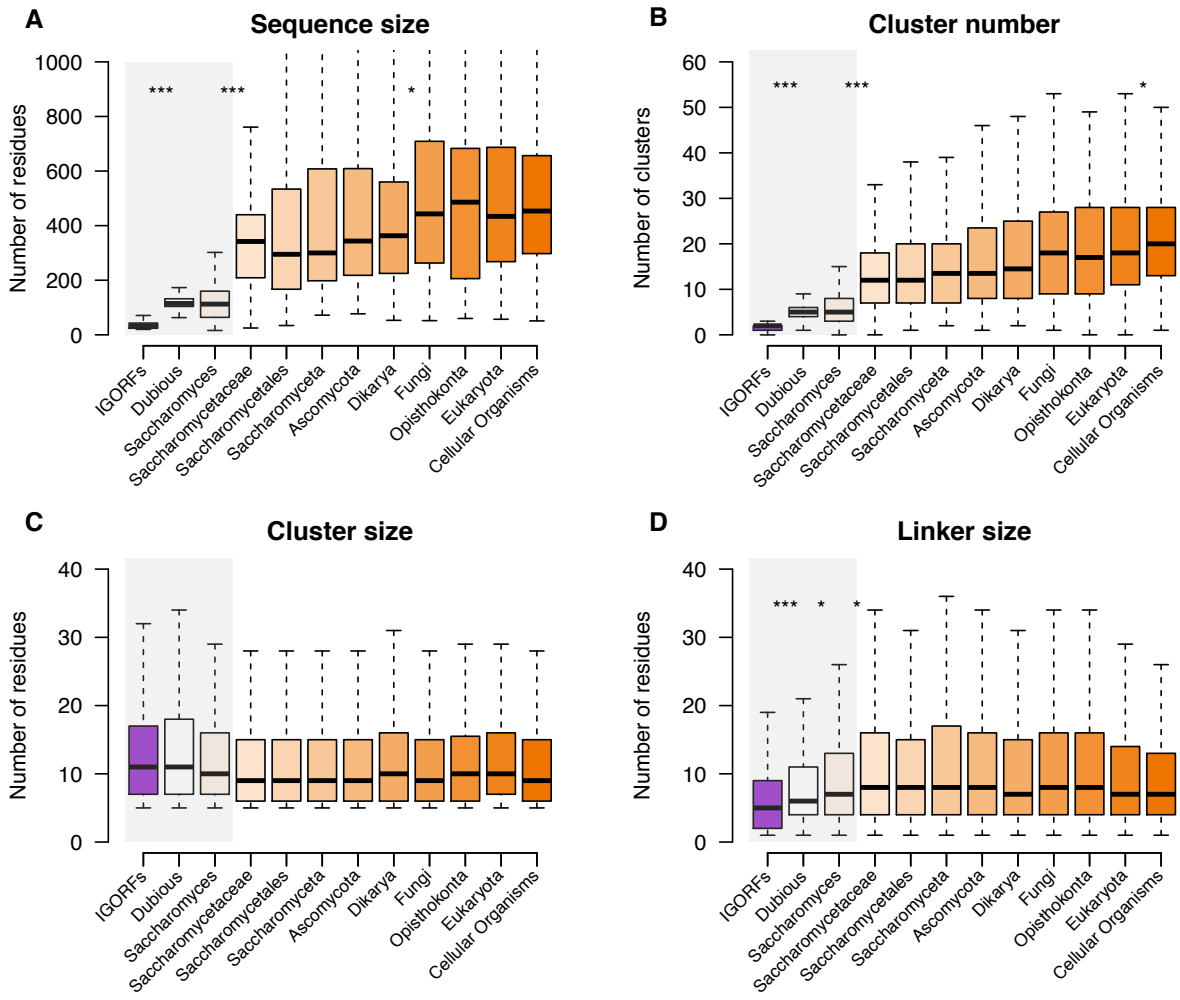


Figure 4.2. Boxplot distribution of the sequence size **(A)**, the number of HCA clusters **(B)**, the size of HCA clusters **(C)** and the size of linkers **(D)** per protein age group (presented in different shades of orange) together with the ones of the IGORFs (presented in purple) and the dubious genes (presented in white). Asterisks denote level of significance for the Mann-Whitney *U*-test for every consecutive pair: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$. The outliers of the boxplots are omitted for clarity.

4.3.3 Evolution of the amino acid composition

As it has already been described in section 3, IGORFs and CDS present distinct profiles of amino acid frequencies. Precisely, IGORFs are mostly enriched in strong hydrophobic amino acids (F, L, I, Y, C) all participating in the HCA clusters while CDS are mostly enriched in polar and charged residues with an even stronger enrichment in negatively charged amino acids (D, E). Figure 4.3 shows the frequency of all 20 amino acids per gene age group. All the age groups, except of the young Saccharomyces TRGs (in beige), present similar amino acids' frequency profiles to the one of CDS (orange thick line). This is not the case for the youngest proteins (in beige) which present a profile intermediate between IGORFs and the older yeast proteins for most amino acids. Notably, the dubious genes (in black) present amino acids frequencies similar to the ones of IGORFs and de novo genes (in blue).

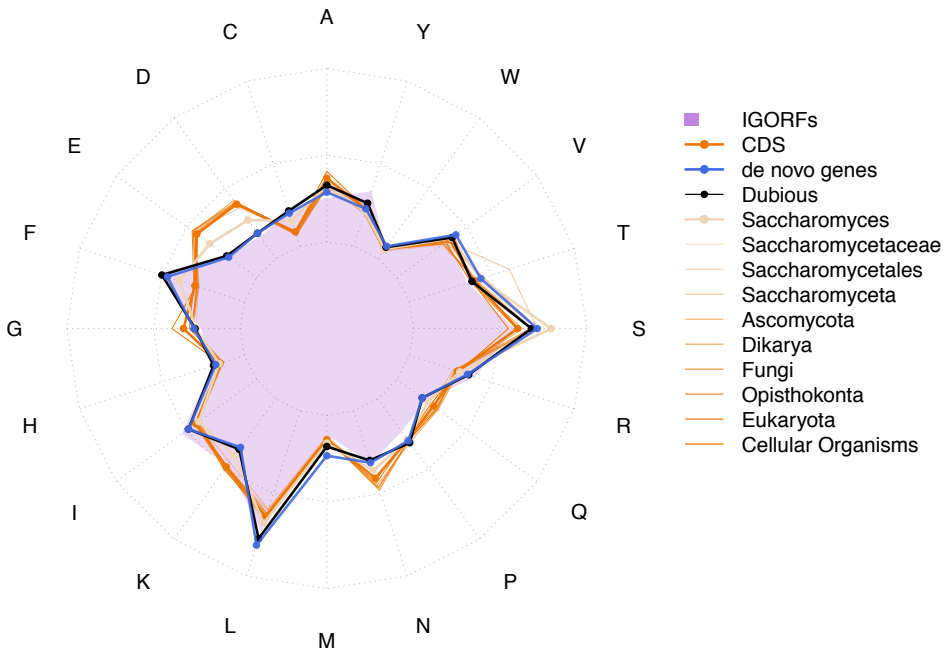


Figure 4.3. Radar plot reflecting the 20 amino acid frequencies of IGORFs (presented in purple) and the CDS for the 10 age groups (presented in different shades of orange) together with the dubious genes (presented in black) and the 70 de novo genes (presented in blue) presented at section 3 of the manuscript.

Figure 4.4A represents the frequencies of amino acids, this time grouped according to their physico-chemical properties, across the different protein phylostrata. Strong hydrophobic amino acids (usually associated with HCA clusters) are represented with empty circles, polar hydrophilic amino acids (usually associated with HCA linkers) are represented with black circles, negatively charged residues (D, E) are shown with red inverted triangles while positively charged residues (K, R) are shown with green triangles. On the plot, we observe that the young genes (Saccharomyces group) present intermediate frequencies of hydrophobic and hydrophilic residues between IGORFs and the older genes. In fact, the young genes are depleted in hydrophobic residues (Mann-Whitney *U*-test, $P = 4 \times 10^{-2}$) and enriched in hydrophilic and specifically negatively charged residues (Mann-Whitney *U*-test, $P = 2 \times 10^{-2}$ and $P = 3 \times 10^{-2}$, respectively) when compared with the IGORFs. On the contrary, no significant difference is observed when they are compared with all the older age groups (Mann-Whitney *U*-test, all pairwise p-values $> 5 \times 10^{-2}$ for hydrophobic, hydrophilic and negatively charged residues, respectively). Consequently, these results support that the hydrophobic and hydrophilic residues frequencies of young yeast genes are more similar to the ones of the older genes highlighting the importance of the negative charges in the genes' evolution. On the other hand, dubious genes were enriched in hydrophobic (Mann-Whitney *U*-test, all pairwise p-values $< 2 \times 10^{-16}$) and depleted in hydrophilic and negatively charged residues (Mann-Whitney *U*-test, all pairwise p-values $< 2 \times 10^{-16}$ and 2×10^{-4}, respectively) compared with all the other age groups while they did not present significant difference of hydrophobic and negatively charged residues when compared with the IGORFs (Mann-Whitney *U*-test, $P = 5 \times 10^{-1}$ and $P = 5 \times 10^{-1}$, respectively) thus reflecting mostly IGORF-like characteristics. Whether dubious genes correspond to an intermediate state between noncoding sequences and young genes (resembling mostly to IGORFs) and whether the young genes are also an intermediate state towards the establishment of older genes are interesting questions.

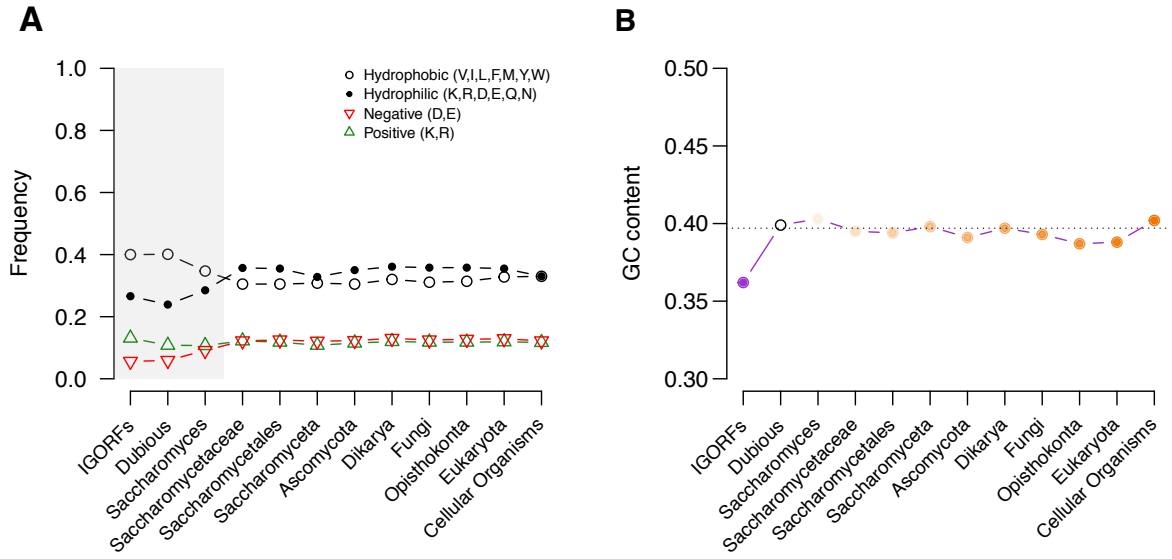


Figure 4.4. (A) Frequency of the strong hydrophobic (presented in empty circles), the hydrophilic (presented in black circles), the negatively charged (presented in red inverted triangles) and the positively charged (presented in green triangles) residues for the different age groups together with the IGORFs and the dubious genes **(B)** GC content of the *S. cerevisiae* CDS for the different age groups together with the IGORFs and the dubious genes. In dotted line the average GC content of the total CDS.

At note, the yeast genes of all the age groups as well as dubious genes, present similar GC content of around 40% while the noncoding ORFs present lower CG content of 36.1% (Figure 4.4B). However, proteins encoded by dubious genes present amino acid composition similar to the one of IGORFs while the young proteins in Saccharomyces group present an overall intermediate composition of amino acids between IGORFs and older genes, thus resembling mostly to CDS when amino acids are grouped according to their physico-chemical properties. In line with section 3 and Vakirlis et al. (2018), is again tempting to speculate that the increased GC content is an important trait for the emergence of novel longer ORFs (stop codons are AT-rich) from noncoding regions and then follows the amino acid optimization by mutating towards negatively charged residues.

4.3.4 Evolution of cellular abundance and number of protein interactions

Our results highlight that the yeast genes in the youngest age group present intermediate characteristics (HCA foldability scores, disorder and aggregation propensity and linker

sizes) between nongenic sequences and older genes. In this part, we are interested in further estimating the potential behavior of a protein in the cell. Therefore, we investigated the relationship between the protein ages and their cellular abundance as well as their number of Protein-Protein Interactions (PPI).

In Figure 4.5A is presented the protein cellular abundance distribution (in parts per million) for the different yeast protein ages together with the dubious genes. Notably, the abundance of the proteins in the cell increases continuously with their age, supporting that the older proteins are more expressed or potentially longer living in the cellular environment. In addition, the proteins in the youngest group together with the dubious genes present significantly lower cellular abundance than any other age group. Indeed, is known that de novo emerging genes present lower expression levels.

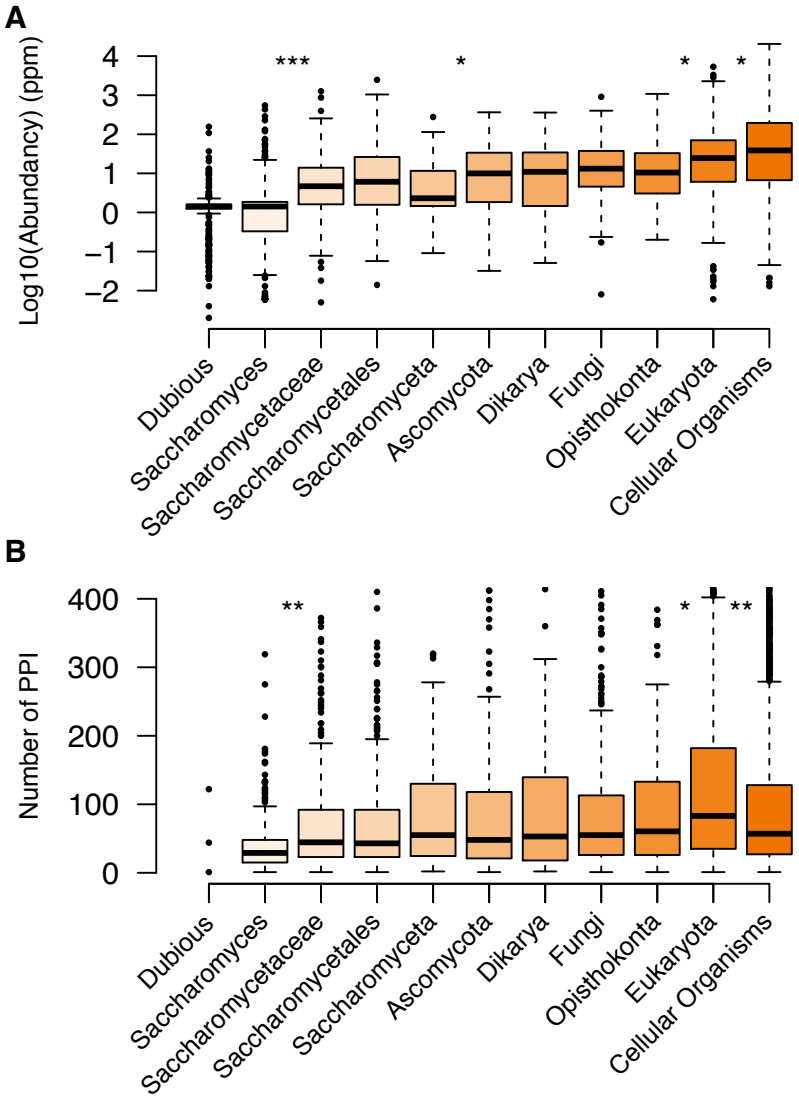


Figure 4.5. Boxplot distribution of **(A)** the protein cellular abundance (in parts per million) and **(B)** the Protein-Protein Interaction partners number per age group (presented in different shades of orange). Asterisks denote level of significance for the Mann-Whitney *U*-test for every consecutive pair: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$.

The study of Abrusán (2013) inspired us to study the evolution of the number of Protein-Protein Interactions (PPI) among the different phylostrata. To do so, we extracted all the yeast PPIs from the BioGRID (version 4.4.200) database (Stark et al. 2006). Figure 4.5B presents the number of PPIs for every yeast protein by age group. In line with Abrusán (2013), it is observed that the young genes' group present fewer PPIs compared with the rest of the age groups (Mann-Whitney *U*-test, all p-values $< 5 \times 10^{-2}$). In addition, even though the consecutive pairwise comparisons between the age groups do not support any significant gradual increase of the PPIs with the evolution time, the comparison between the younger and the older phylostrata revealed that the three younger phylostrata (from Saccharomyces until Saccharomycetales group) contain significantly less PPI partners than the four oldest age groups (from Fungi until Cellular Organisms group). Whether younger proteins interact with fewer partners due to more specific cell functionalities is a question which should be further investigated. In addition, it must be mentioned that as the interaction databases are incomplete, the number of PPIs is likely to be underestimated and consequently these results should be considered with cautiousness.

We showed that the protein cellular abundance and the number of PPIs, both present an increasing tendency with the protein age. In addition, the young yeast proteins of *S. cerevisiae* presented significant lower abundance and less PPIs than any other age group. Whether these two observations reflect an intermediate character of young proteins which are less expressed and interact with fewer partners than their older counterparts, is an interesting question. However, it must be highlighted that although the two properties studied in this section (cellular abundancy and number of PPIs) present similar increasing tendencies along evolution time, we do not know if there is any correlation between them and further studies should be done to investigate that.

4.3.5 Evolution of protein fold

Our results based on HCA, highlight that the proteins' fold potential is continuously optimized with the evolutionary time. The HCA score provides information about the capacity of a protein to fold but does not provide any information about its structural content. In this part, we use the proteins' age groups of the yeast in order to investigate if and how the evolution has shaped the overall structural content of proteins.

4.3.5.1 Prediction of structural domains from the protein sequences

Different classification methods have been developed in order to organize and categorize the folds' universe (Hubbard et al. 1997; Andreeva et al. 2014; Dawson et al. 2017; Sillitoe et al. 2021; Mistry et al. 2021). One of these manual classification schemes is the SCOP database which attempts to cluster hierarchically protein domains with common structural and evolutionary relationships (Hubbard et al. 1997; Andreeva et al. 2014). Based on protein domains' evolutionary divergence and structural similarity, SCOP organizes them into families and superfamilies. These are further classified into structural folds, which do not necessarily indicate common evolutionary origin, and classes reflecting the domains' secondary structures (Andreeva et al. 2014). Globular proteins' domains are classified by their majority secondary structure content in one of the four main SCOP classes (Hubbard et al. 1997; Edwards et al. 2013):

1. **all-α**: those whose structure is essentially formed by α-helices
2. **all-β:** those whose structure is essentially formed by β-sheets
3. **α/β:** those with α-helices and β-strands
4. **α+β:** those in which α-helices and β-strands are largely segregated

or the fifth class:

5. **multi-domain:** those with domains of different fold and for which no homologues are known at present.

In this part, we annotated the domain superfamilies for the proteins of *S. cerevisiae* in order to investigate the domains' representation in different age groups. SUPERFAMILY is a library of hidden Markov models (HMMs) for sequences corresponding to protein domains with known structure (Gough et al. 2001). These HMMs are used in order to identify SCOP superfamilies of domains in protein sequences (Gough et al. 2001). SUPERFAMILY model library has been used in order to annotate the sequences of over

50 proteomes, one of which was the proteome of *S. cerevisiae*. We retrieved the domains' annotation for the proteome of *S. cerevisiae* and assigned every domain with its corresponding SCOP class (all-α, all-β, α/β, α+β, multi-domain) together with the age group of the protein the domain belonged. It is important to mention that there are proteins of *S. cerevisiae* with no domain annotation (~45% of the proteome) while others with more than one domain. In Table 4.2 are presented the counting of proteins with at least one annotated domain as well as the total number of domains annotated per age group. Is interesting to highlight that 90.8% (2339/2575) of the oldest proteins were annotated with at least one SCOP superfamily while it is the case only for 1.4% (8/562) of the youngest ones. Notably, the oldest age group is overrepresented (one proportion z-test, $P = 1 \times 10^{-8}$) while the three younger age groups (from Saccharomycetales to Saccharomyces) are underrepresented (one proportion z-test, $P = 1 \times 10^{-2}$, $P = 6 \times 10^{-3}$ and $P = 2 \times 10^{-9}$, respectively). In addition, dubious genes did not present any domain annotation.

Table 4.2. Count of *S. cerevisiae* proteins with at least one annotated domain and total number of annotated domains per age group.

	Total number of proteins	**Proteins with at least one domain**	**Total number of domains**
Cellular Organisms	2575	2339	3634
Eukaryota	1578	989	1372
Opisthokonta	127	49	59
Fungi	289	115	128
Dikarya	75	18	20
Ascomycota	140	33	37
Saccharomyceta	71	10	10
Saccharomycetales	366	34	36
Saccharomycetaceae	346	16	18
Saccharomyces	562	8	9
Dubious	545	0	0
Total	6649	3611	5323

In Figure 4.6 is presented the repartition of the 5323 domain classes annotated for 3611 proteins of *S. cerevisiae* among the different phylostrata. Due to the low number of annotated domains for the younger proteins, we decided to pull together some neighboring age groups in order to have a more statistically accurate representation (the groups are highlighted with dotted lines in Table 4.2). It is interesting to note that proteins of the oldest age group are significantly enriched in α/β domains (Mann-Whitney *U*-test, all p-values < 2.2×10^{-16}) and significantly depleted in all-α domains (Mann-Whitney *U*-test, all p-values < 2.2×10^{-3}) compared with any other age group, supporting that older proteins have evolved towards domains combining alpha helix and beta strand secondary structures while younger proteins mostly contain domains with similar secondary structures (all-α or all-β).

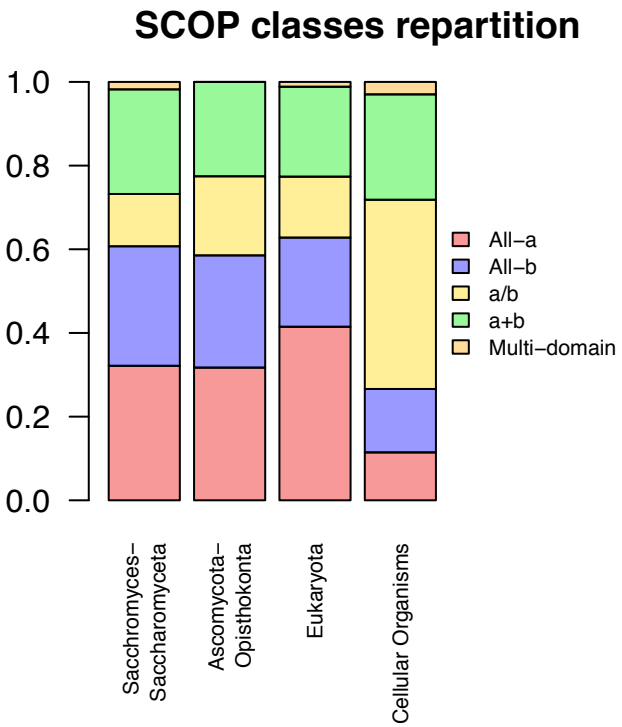


Figure 4.6. Repartition of the annotated SCOP classes for protein domains of different age groups.

4.3.5.2 Structural content of *S. cerevisiae* proteins

Based on the SCOP superfamilies annotation, we observe that domains in older proteins tend to present a mixed composition of secondary structures while the domains of

younger proteins tend to be more homogenous. In this part, we investigate whether the same observations stand for the 3D structures of *S. cerevisiae* proteins.

We were able to extract up to 1346 protein structures from the PDB, each one corresponding to a single yeast gene (see Methods for details) and assign them with one of the ten phylostrata. Although these proteins cover only almost 20% of the total *S. cerevisiae* proteome, the advantage of this dataset is that it contains structures that were characterized experimentally. On the other hand, in order to increase the coverage of the yeast proteome, we used the 3D protein structural models as predicted by AlphaFold2 for the total proteome of *S. cerevisiae*. Recently published AlphaFold2, is a method that relies on deep learning for predicting the 3D structure of a protein from its amino acid sequence (Jumper et al. 2021; Tunyasuvunakool et al. 2021). Contrary to the most successful free modelling approaches which rely on fragment assembly to predict the shape of a protein, AlphaFold2 is trained on PDB structures in order to predict the pairwise distances between the Cβ atoms of a protein's residues. Distance predictions provide more specific information about the shape of the protein than contact predictions. With this information, AlphaFold2 constructs a potential of mean force that can accurately describe the shape of a protein, and which can further be optimized in order to generate more accurate structure predictions (Senior et al. 2020). The central component of AlphaFold2 is a convolutional neural network trained on PDB structures and has been shown to achieve high accuracy, even for sequences without a template structure in the PDB or with relatively few homologous sequences (Senior et al. 2020). Recently, the AlphaFold team together with the European Molecular Biology Laboratory (EMBL) released a publicly available database which contains protein structure predictions, made with AlphaFold2, for the whole proteome of many different model organisms (i.e., human or yeast) (Tunyasuvunakool et al. 2021). We extracted up to 5974 structure predictions covering almost 90% of the *S. cerevisiae* proteome and assigned them with one of the age groups.

In Table 4.3 is presented the repartition of the experimental protein structures and the predicted structural models of *S. cerevisiae* in the ten phylostrata. Concerning the 1346 PDB protein structures of *S. cerevisiae*, it should be highlighted that the two older age groups (Cellular Organisms and Eukaryota) are overrepresented by structures (one

proportion z-test, $P = 1 \times 10^{-3}$ and 3×10^{-3}, respectively) while the younger age group is significantly depleted in structures (one proportion z-test, $P = 1 \times 10^{-9}$) reflecting the unequal representation of the different age groups in the PDB. On the contrary, all the phylostrata are well represented by AlphaFold[2] models (one proportion z-test, all p-values $> 5 \times 10^{-2}$) with only exception the dubious genes which are underrepresented (one proportion z-test, $P = 5 \times 10^{-3}$).

Table 4.3. Count of *S. cerevisiae* proteins with a 3D structure in the PDB and a 3D protein structure model predicted by AlphaFold[2] per age group.

	Total number of proteins	**Proteins with PDB structures**	**AlphaFold models**
Cellular Organisms	2575	720	2555
Eukaryota	1578	428	1553
Opisthokonta	127	23	125
Fungi	289	39	289
Dikarya	75	9	74
Ascomycota	140	21	140
Saccharomyceta	71	14	70
Saccharomycetales	366	44	351
Saccharomycetaceae	346	38	343
Saccharomyces	562	10	360
Dubious	545	0	114
Total	6649	1346	5974

AlphaFold[2] calculates a confidence metric per-residue (on a scale from 0 to 100) for every prediction, called predicted lDDT-Cα (pLDDT). The confidence score estimates how well the prediction is expected to agree with an experimental structure based on the Local Distance Difference Test (Tunyasuvunakool et al. 2021). A value of pLDDT > 90 is considered as the high accuracy cutoff, corresponding to a correct prediction at the level of residue's side chain. A lower cutoff of pLDDT > 70 corresponds to a generally correct backbone prediction while low pLDDT values (lower than 50) should not be interpreted

as structures but rather as a prediction of disorder (Tunyasuvunakool et al. 2021). As a result, AlphaFold[2] defines four distinct confidence score thresholds:

- pLDDT > 90 : High accuracy prediction – Correct side chain
- pLDDT between 70 and 90 : Confident prediction – Correct backbone
- pLDDT between 50 and 70 : Low accuracy prediction
- pLDDT < 50 : Very low accuracy prediction – Disordered region

In Figure 4.7A is presented the median value of the frequency for high (in blue), confident (in light blue), low (in orange) and very low (in yellow) accuracy predicted amino acids per age group and the dubious genes. Similarly, in Figure 4.7B is presented the same information but this time the high and confident accuracy predictions are summed together (in blue) and the low and very low accuracy predictions are summed as well (in orange). Strikingly, we observe a clear increasing tendency of high accuracy predicted amino acids with the evolution time while the opposite is true for the low confident predicted ones.

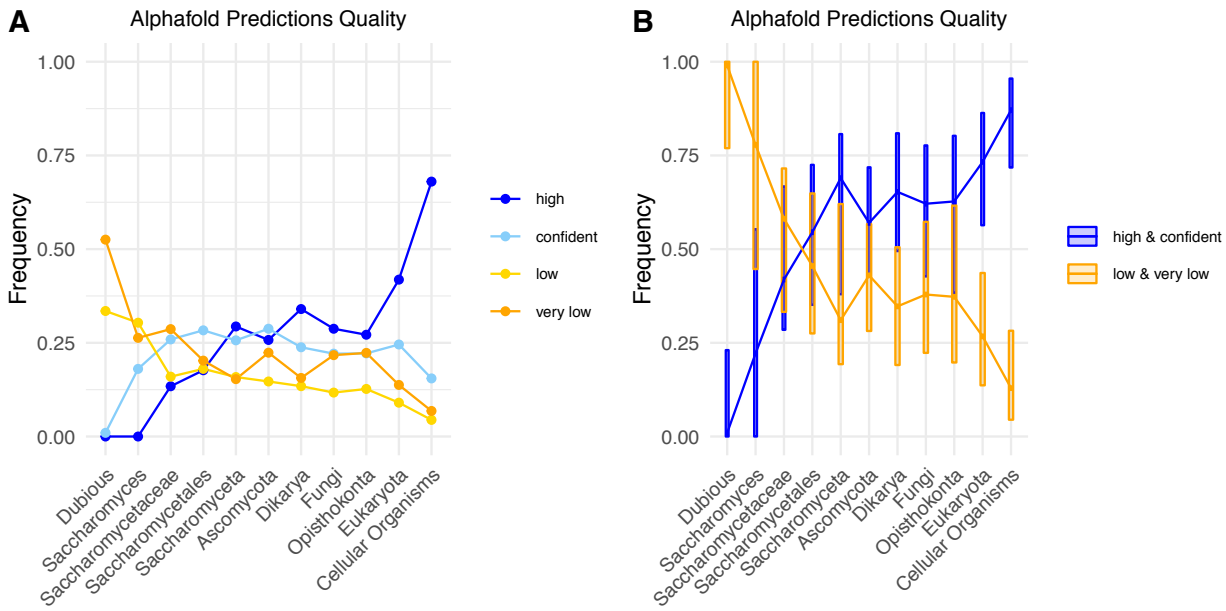


Figure 4.7. (A) Frequency of high (in blue), confident (in light blue), low (in orange) and very low (in yellow) accuracy predicted amino acids per age group **(B)** Frequency of high and confident accuracy predictions together (in blue) and the low and very low accuracy predictions together (in orange).

These results highlight that AlphaFold² struggles to predict with accuracy the 3D structure of young proteins. This may result from the fact that young genes display less homologous sequences, thus producing less reliable Multiple Sequence Alignments which may lead to low accuracy scores. This may also result from the fact that young genes encode proteins with folds absent from the PDB, on which AlphaFold² has been trained, in line with the observation that young proteins are underrepresented in the PDB dataset of the *S. cerevisiae* proteome. Additionally, this may reflect the fact that young genes encode proteins with different foldability properties like the rudimentary fold proposed for the young protein Bsc4 which is folded but lacks a specific and well-defined 3D fold. Whether young proteins are well folded or display a rudimentary fold deserves further investigation. The wide range of the HCA score for the young yeast proteins could support this hypothesis although this must be demonstrated. As an example, the prediction of AlphaFold² on the Bsc4 protein (Figure 4.8) presented an overall rudimentary fold with few secondary structures (57% of the residues in coil conformation) and very low prediction score. Precisely, 13.7% of the residues presented low prediction score while the remaining 86.3% presented very low score.

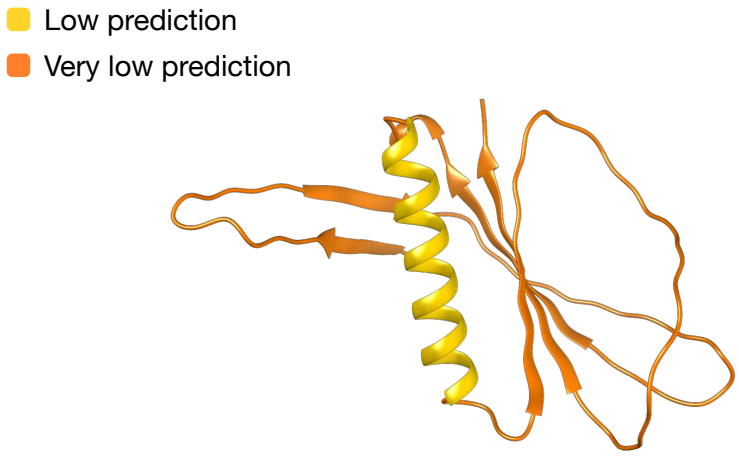


Figure 4.8. Prediction of the 3D structure of the Bsc4 protein made by AlphaFold². In yellow the residues with low prediction confidence and in orange the residues with very low prediction confidence.

We used the stride tool in order to assign the secondary structure for each protein of our two datasets (PDB structures and AlphaFold² models). Based on local geometrical arrangements of atoms, stride can assign a secondary structure state (alpha helix or beta

strand) or coil state to each amino acid of the protein. Specifically, for the AlphaFold[2] models we assigned the secondary structures for all the residues predicted with pLDDT score more than 50 in order to avoid residues with very low quality of prediction. Then for every protein we calculated the fraction of residues assigned in alpha helix and in beta strand secondary structure. In order to estimate the secondary structure composition of a protein we calculated the absolute value of the difference of these two fractions:

$$\text{Secondary Structure Composition} = \left| \frac{\text{residues}_{\text{alpha}}}{\text{residues}} - \frac{\text{residues}_{\text{beta}}}{\text{residues}} \right|$$

Low values of this metric correspond to proteins containing important fractions of both alpha helices and beta strands while higher values correspond to proteins mostly populated by one of the two secondary structure types. In Figure 4.9 is presented the secondary structure composition for the proteins of the PDB (Figure 4.9A) and the AlphaFold[2] models (Figure 4.9B).

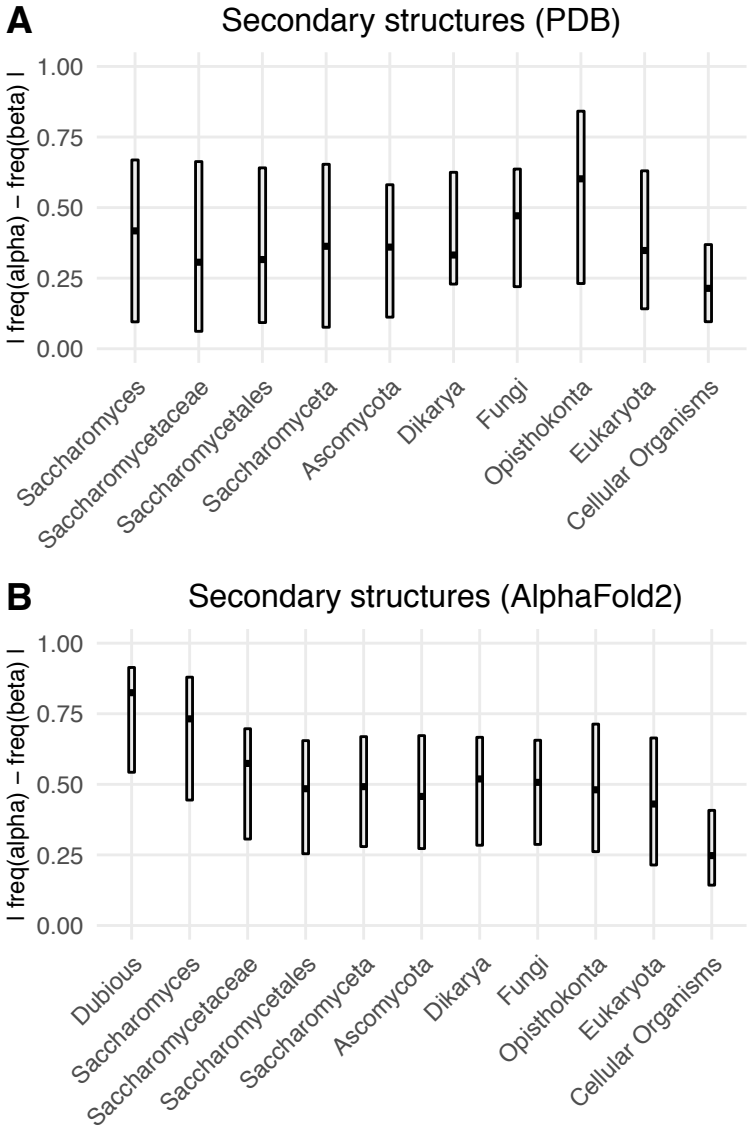


Figure 4.9. Boxplot distribution of the difference between the fraction of residues in alpha helix and the fraction of residues in beta strand by age group for **(A)** the experimental protein structures of the PDB and **(B)** the protein models predicted by AlphaFold[2].

In figure 4.9B (AlphaFold[2]) we can observe a decreasing tendency of this metric from younger towards older proteins reflecting that the proteins evolve towards structures with more mixed composition of secondary structures combining alpha helices and beta strands while younger proteins tend to present a more homogenous representation of secondary structures (mostly all alpha or all beta). The tendency is less pronounced for the dataset of PDB proteins (Figure 4.9A) but this can be associated with the low representativity of the proteins in the PDB subset. Nevertheless, the older proteins

present significantly lower values of this metric compared with the proteins of any other age group (Mann-Whitney U-test, all p-values < 2.2×10^{-7}). These results are in line with the results obtained for the protein domains by the SCOP superfamilies annotation.

4.4 Conclusions

In this section we were interested in investigating if and how numerous structural properties of *S. cerevisiae* proteins variate along with the evolutionary time and consequently understand how evolution has shaped the structural properties of proteins. Our results on the proteins' fold potential highlight that the yeast proteome tends to become more foldable with the evolutionary time, supporting that foldability is an important feature that gets continuously optimized with evolution. As a matter of fact, the example of young de novo protein Bsc4, with its rudimentary fold, supports that young proteins present structures which lack fold specificity compared to older well folded proteins. Globular proteins are usually characterized by a stable and well folded structure known to be a requirement for many aspects of their function (Edwards et al. 2013). As a result, an interesting question is whether the fold potential optimization observed for older proteins, is intimately related with a potential functional optimization or whether is the outcome of the selection pressure exerted on them for longer time. Our results do not permit us to conclude.

Additionally, we observed the well reported increasing tendency of protein sequence size accompanied by the increase of HCA clusters occurrences along with the evolutionary time. Interestingly, the HCA cluster sizes for all the protein age groups (including the dubious and the IGORFs) present similar distributions across the different protein ages, enforcing even more our initial results and giving to the HCA clusters an interesting role as elementary protein building blocks of constant sizes throughout the evolutionary time. On the other side, the HCA linkers of the CDS presented also similar sizes among the yeast phylostrata, with exception the youngest age group (Saccharomyces TRGs) which present intermediate linker sizes between IGORFs and older proteins showing that size of linkers is a property that is fixed early in protein evolution. This result is strongly supported by the disorder and aggregation propensity swifts observed between the young Saccharomyces genes and the young established Saccharomycetaceae genes.

Whether this swift of young proteins towards sequences with longer disorder regions and less aggregation propensity is an internal criterion for the foldability optimization along evolution or is simply the outcome of the sequences' elongation, is an interesting question that we pose. Edwards et al. (2013) compared newly evolved structures with structures presenting a long evolutionary history and showed that, overall, a shorter evolutionary history corresponded to less elaborate structures with fewer intra-residues contacts. They speculated that newly born proteins evolve into gradually longer ones maintaining at the same time the positioning of important residues while minimizing other undesirable interactions. In fact, study of their amino acid frequencies showed that young yeast proteins present frequencies of hydrophobic and hydrophilic residues similar to the ones of their older counterparts, although the effect is marginal, while proteins encoded by dubious genes presented amino acid frequencies similar to IGORFs.

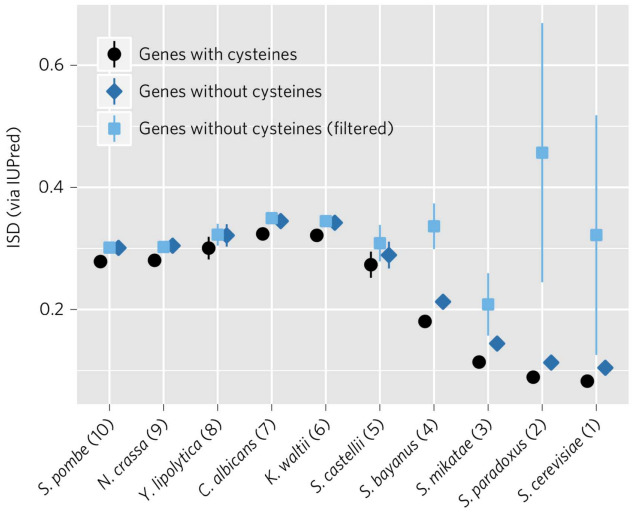
Studying the repartition of the per-residue confidence score of AlphaFold^2 along the different phylostrata but also the dubious genes, we observed that AlphaFold^2 structural models for young and dubious proteins present very low confidence scores while the score increases along with the evolutionary time. This result highlight that AlphaFold^2 struggles to predict with high confidence the structure of young proteins or proteins encoded by dubious genes. This could be related to the lack of homologous sequences for young proteins but also permits us to speculate that these proteins present a more rudimentary fold which is less represented in the PDB. Notably, proteins encoded by dubious genes presented an intermediate character of structural and sequence properties between IGORFs and young *S. cerevisiae* TRGs for some features and young genes-like for others. Whether the dubious genes reflect an intermediate state between noncoding sequences and young yeast genes (reminding the proto-genes of Carvunis et al. (2012)) and whether young genes are also an intermediate state towards older well-established genes in an evolutionary continuum are interesting questions. Regarding this question, Paul Roginski (1st year of PhD), during his Master 2 internship developed a machine learning model which aims at discriminating coding sequences from noncoding ones by training on random nucleotide sequences of similar sizes. His model uses descriptors such as nucleotide and codon frequencies as well as the frequency of the four bases in every codon position and presents high accuracy values. Indeed, his model presented high predictive capacity (95%) on the genes of *S. cerevisiae*. Interestingly, the

majority of the 336 non-predicted genes corresponded to young yeast TRGs (23.2%) as well as dubious genes (44%), supporting that these genes present premature and not yet optimized nucleotide sequences. All these observations remind the continuum hypothesis of Carvunis et al. (2012) which propose the proto-genes as intermediate reversible states of young sequences with intermediate structural properties. These young and weakly expressed sequences, in the absence of selection for some beneficial properties, could easily turn back to their noncoding state while they could be established as young de novo genes in the opposite scenario.

Moreover, although we used the same phylostratigaphy data with Wilson et al. (2017), we did not make the same observations. Firstly, the study of Wilson et al. (2017) was basically focused on the mouse proteome and consequently all their observations were mostly valid for mammalian multicellular species. They referred to the yeast proteome only in order to compare the preadaptation theory with the continuum model presented by Carvunis et al. (2012). In order to compare their results with the ones presented by Carvunis et al. (2012), they fitted their data to the phylostratigraphy made by the proto-genes study, which did not present the same age groups delimitations. Then, they presented the disorder propensity of the yeast proteins based on the age groups as defined by Carvunis et al. (2012) (Figure 4.10A). Interestingly, they observed the same tendency as described by Carvunis et al. (2012), that young proteins tend to present lower disorder propensity than older proteins (Figure 4.10A black circles and blue diamonds). Then, they applied some filters in order to curate their data. Notably, they removed all the dubious genes, the genes that they were not able to classify in a clear phylostratum and the genes that did not present any homolog except of *S. cerevisiae* (*S. cerevisiae* orphan genes), considering them as potential annotation errors, thereby eliminating all recent de novo genes. The counting of the genes per phylostrata after the application of every filter is presented in figure 4.10B. After their data treatment, the two younger phylostrata (*S. cerevisiae* and *S. paradoxus*), which are precisely the ones that support their theory of higher disorder of young proteins contrarily to what has been observed by Carvunis et al. (2012) (figure 4.10A Light blue squares), contained only 2 genes each (Figure 4.10B – lines 1 and 2, last column). In fact, in figure 4.10A, the standard error bars of the distribution of ISD values for the two younger phylostrata are very long simply because they contain only 2 observations each. This reduces considerably the

confidence one can have in their results. On the contrary, our results clearly agree with those of Carvunis et al. (2012).

A



B

Table 1 | Number of genes assigned to each of the conservation levels annotated in ref. [3].

Conservation level	Genes matching our dataset	After excluding dubious genes	After excluding genes that we found to be unique to *S. cerevisiae*	After excluding genes whose age we were unable to classify
1 (*S. cerevisiae*)	143	35	3	2
2 (*S. paradoxus*)	172	44	2	2
3 (*S. mikatae*)	137	31	4	4
4 (*S. bayanus*)	325	162	28	27
5 (*S. castellii*)	94	88	53	53
6 (*Kluyveromyces waltii*)	511	501	489	489
7 (*C. albicans*)	411	405	398	398
8 (*Yarrowia lipolytica*)	81	79	78	77
9 (*Neurospora crassa*)	499	491	483	482
10 (*S. chizosaccharomyces pombe*)	3,935	3,925	3,918	3,918

Figure 4.10. (A) Prediction of the disorder propensity (ISD) of the *S. cerevisiae* proteins for the different phylostrata as defined by the phylostratigaphy presented in Carvunis et al. (2012). At note, the direction of the phylostrata is from the older towards the younger. Black circles and blue diamonds correspond to the ensemble of proteins before applying any filtering to the dataset (including and excluding the cysteines from the protein primary sequence, respectively). Light blue squares correspond to the proteins after excluding dubious genes, *S. cerevisiae* orphan genes and unclassified genes based on the study of Wilson et al. (2017). **(B)** The counting of the proteins per phylostratum after excluding dubious genes, *S.cerevisiae* orphan genes and unclassified genes. Both, the figure and the table were extracted from the study of Wilson et al. (2017) with title *"Young genes are highly disordered as predicted by the preadaptation hypothesis of* de novo *gene birth"*.

Concerning the structural content of the yeast proteome, our results highlight that young proteins tend to present a more homogeneous secondary structure content (all-α or all-β) while, on the contrary, older proteins seem to acquire folds that combine alpha and beta secondary structures thus leading to more complex arrangements such as the α/β fold class. Our results are in line with Choi and Kim (2006) who showed that recently born and still-evolving proteins belong mostly to all-α or all-β class (as well as their random mixtures, α+β class), while the majority of the older aged proteins belong to α/β class. Notably, Edwards et al. (2013) analyzed all the SCOP superfamilies and showed that

α/β class domains were significantly older than superfamilies belonging to the other three classes. Interestingly, α/β folds also contain a large number of the so-called 'superfolds' (folds containing large numbers of different superfamilies and a high proportion of all determined structures) which are known to be associated with a large repertoire of fundamental functions. Enrichment analysis of GO functions in the different age groups demonstrated that ancient superfamilies correspond to fundamental cellular processes shared among the vast majority of the species (Edwards et al. 2013). Choi and Kim (2006) proposed a scenario for the evolution of the protein structural classes which supports that the ancestral proteins contain mostly short secondary structures and consequently correspond to three SCOP classes (all-α, all-β and α+β). Then they evolve to medium-sized proteins and are distributed in the four SCOP classes (all-α, all-β, α+β and α/β) while finally they evolve to large proteins populating mostly the α/β class. Alva et al. (2010) in their study "A galaxy of folds" generated a network of all the SCOP fold classes connected according to their sequence similarity. They observed that even though the fold classes were in general well clustered, there were numerous incidences of domains from different superfamilies and folds with homologous connections, thus supporting the potential of interchange among different fold types during evolution. Whether the combination the two types of secondary structures, observed for the older proteins, is related with more diverse and fundamental functions or is simply the outcome of the proteins' size increase which permit more complex rearrangements is a question to investigate.

5 Prediction of the fold state of peptides using machine learning

5.1 Introduction

In section 3, it was presented a large-scale analysis of the foldability potential for the ensemble of *S. cerevisiae* IGORFs. Using the HCA method, we were able at defining and predicting three major categories of sequences: (i) Sequences with low HCA-score, rich in polar and charged amino acids that potentially encode highly disordered peptides. (ii) Sequences with high HCA-score, rich in strong hydrophobic amino acids potentially encoding transmembrane peptides and believed to aggregate in solution due to high exposure of hydrophobic residues. (iii) Sequences with intermediate HCA-score presenting a mixture of hydrophobic and hydrophilic amino acids, that potentially encode peptides expected to be able to acquire a 3D fold under solution or upon oligomerization or interaction with other protein partners. However, it must be mentioned that folding into a well-defined conformation is a stricter requirement than forming a molten globule with some secondary structure elements (Mezei 2020). LaBean et al. (2011) showed that unevolved random polypeptides fold mostly into a molten globule conformation. This reveals that the foldable category contains an important variability of fold potentials ranging from simply forming some secondary structures arranged around a hydrophobic core (rudimentary fold like the example of Bsc4 protein) until being well folded like a globular protein. More generally, the concept of protein foldability should be systematically used in a specific context. Indeed, the case of the transmembrane helices highlights the importance of the context under which a peptide can acquire its fold state. For instance, the TM (i.e., Transmembrane) peptides are folded in a hydrophobic environment (i.e., the membrane) but probably misfolded and expected to aggregate in solution. Here, the foldable category of HCA concerns the capacity of a protein of acquiring a fold in solution (i.e., in the cytosol).

In this chapter we aim at better characterizing the foldability potential and finally the conditions necessary to the potential peptides encoded by IGORFs to fold. In particular,

we wish to better discriminate those that adopt a stable fold in solution, from those that display a more rudimentary fold such as the de novo protein Bsc4 or those that are disordered in solution but able to fold upon binding with a partner. This will indirectly provide us with information about their potential behavior in the cell and more generally will offer an opportunity to further study the structural properties of specific ORF categories (i.e., whole population of IGORFs, occasionally and highly translated IGORFs) thereby investigating whether the different ORF categories display specific structural properties and foldability status.

If a lot of successful methods have been developed to predict the 3D structure of a protein, they are not designed to distinguish peptides with a stable fold in solution from disordered ones which fold upon binding or peptides with a rudimentary fold. In fact, they were mostly trained on well-defined and stable 3D structures extracted from the PDB, and we can expect that they will struggle to predict the structure of proteins with rudimentary folds as observed for Bsc4 and its prediction from AlphaFold[2]. Therefore, in this section we aim to develop a method able to characterize more specifically the different foldability status of the potential peptides encoded by the IGORFs of a genome. The method must be fast enough to handle several thousands of peptides and will focus on the characterization of short amino acid sequences (20 – 70 residues). Indeed, our results presented in section 3 suggest that the sequence length has an effect on the structural properties and finally on the foldability of the corresponding peptide or protein. Consequently, we developed a method dedicated to the characterization of the foldability status of peptides and do not guarantee its applicability to larger proteins. To do so, we defined five distinct fold states, and generated a dataset of peptides populating these categories. We calculated numerous sequence-based physicochemical descriptors and constructed a supervised machine learning (ML) pipeline based on multiple two-class Support Vector Machine (SVM) classifiers. Each classifier was trained to predict specifically a single fold state and the predictions of all the SVMs were combined together for the final fold category decision. We then applied our model to characterize the potential fold states of the *S. cerevisiae* IGORF-encoded peptides as well as the peptides encoded by pervasively translated IGROFs in yeast and peptides resulting from the translation of alternative reading frames of human CDS (Brunet et al. 2019, 2021).

5.2. Methods

5.2.1 Dataset for the construction of the model

For the development of the model, we defined five different categories of fold states and generated a database of peptides with known fold state annotated according to these five categories. The size of the peptides ranged between 20 and 70 amino acids, corresponding to the size range of 95% of the IGORFs. The lower limit is the minimum size of IGORFs while the higher limit corresponds to the average plus two times the standard deviation of the size of the IGORFs. The details for the five fold states are presented in Table 5.1.

Table 5.1. Information about the fold state categories used for our prediction model.

Fold state	Symbol	PDB	Database	Count	Comments
Intrinsically Disordered	IDP	No	DisProt	417	Disordered regions extracted from intrinsically disordered proteins
Disordered Binding Site	DIBS	Yes	DIBS	232	IDPs capable of binding to and folding upon the surface of ordered protein partners
Small proteins	Small	Yes	SCOPe (g)	220	Proteins with little or no secondary structures.
Structure Stable in Solution	S3	Yes	SCOPe (a,b,c,d,e)	206	Small proteins with ordered structure in solution
Transmembrane helix	TM	Yes	PDBTM	305	Transmembrane segments extracted from membrane proteins

- **Intrinsically Disordered Peptides (IDPs):** Correspond to disordered regions of intrinsically disordered proteins extracted from the DisProt database (Piovesan et al. 2017; Hatos et al. 2020). The mmseqs tool (Hauser et al. 2016) was used in

order to remove sequences sharing more than 30% of sequence identity (we retained only one representative sequence).

- **Disordered Binding Site (DIBS):** Correspond to peptides disordered in solution but capable of binding to and folding upon the surface of ordered protein partners. They were extracted from the DIBS database (Schad et al. 2018). The mmseqs tool (Hauser et al. 2016) was used in order to remove sequences sharing more than 30% of sequence identity (we retained only one representative sequence).
- **Small proteins (Small):** Correspond to proteins with little or no secondary structures in solution as monomers. In general, they lack an extensive hydrophobic core, and their secondary structures are small and irregular. Their tertiary structure is usually maintained by disulfide bridges (Cheek et al. 2006). They were extracted from the SCOP database (class g) and correspond to structures of single chain or cases that were clearly mentioned as monomers.
- **Peptides with Structure Stable in Solution (S3):** Correspond to small proteins presenting an ordered structure in solution as monomers. They were extracted from the SCOP database (classes a, b, c, d and e) and correspond to structures of single chain or cases that were clearly mentioned as monomers.
- **Transmembrane helices (TM):** Correspond to Transmembrane segments of membrane proteins extracted from the PDBTM database (Tusnády et al. 2004, 2005; Kozma et al. 2012) and predicted as transmembrane helices by the TMHMM tool (Sonnhammer et al. 1998; Krogh et al. 2001). The mmseqs tool (Hauser et al. 2016) was used in order to remove sequences sharing more than 30% of sequence identity (we retained only one representative sequence).

Concerning the S3 and Small fold categories, due to their limited number of sequences we decided to not remove their redundancy but rather to weight the contribution of each sequence during the training of the model (explained later). Some illustrative examples of the 3D structure of the different fold states (except the IDPs) are presented in Figure 5.1.

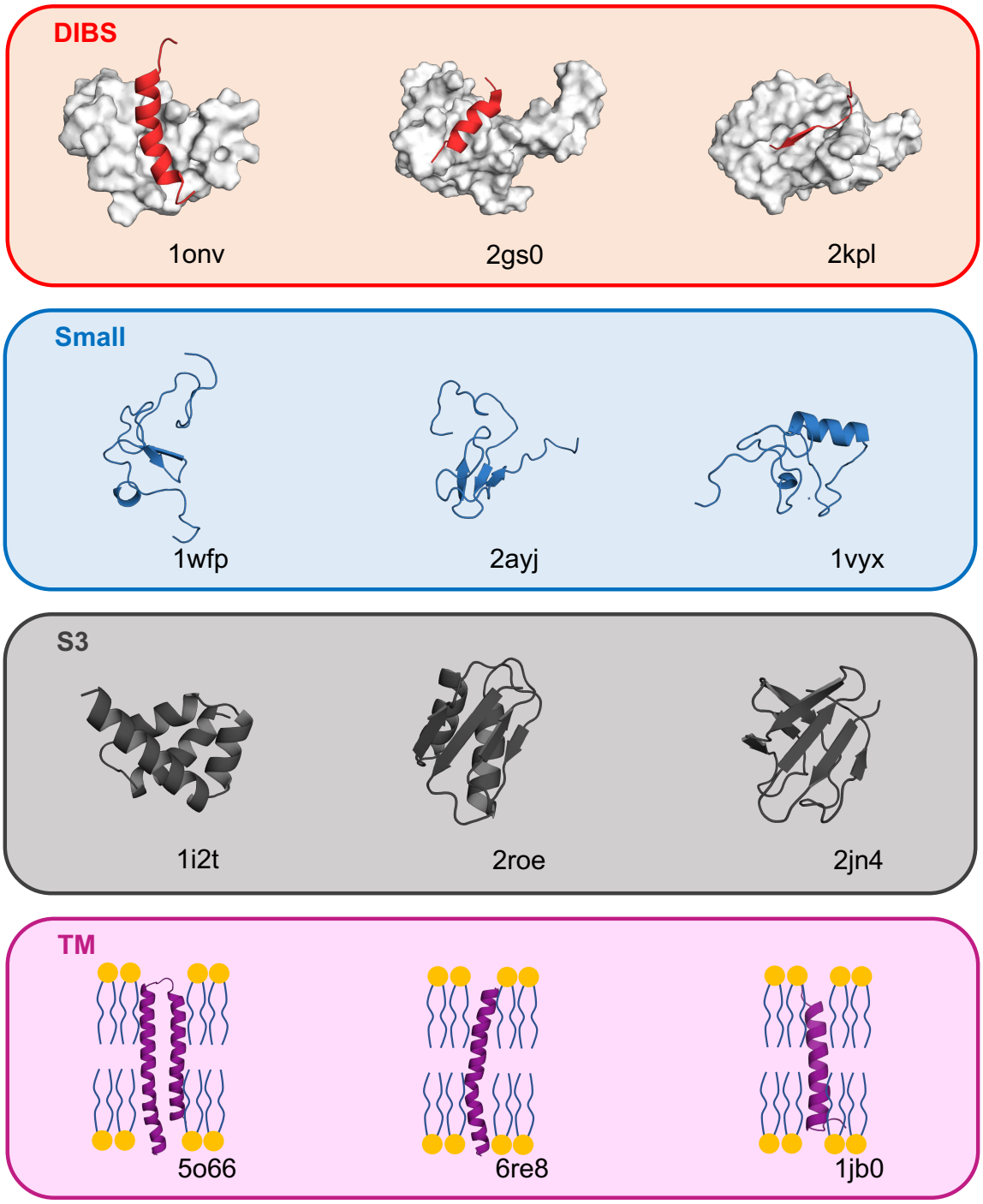

DIBS
1onv
2gs0
2kpl
Small
1wfp
2ayj
1vyx
S3
1i2t
2roe
2jn4
TM
5o66
6re8
1jb0

Figure 5.1. Illustrative examples of some 3D structures for the different fold states (except the IDPs) used for the training of the model. The corresponding PDB codes are mentioned under every peptide. For the TM representation the lipid bilayer is represented schematically.

5.2.2 Datasets for the application of our method

Once the model was constructed, we applied our method on different categories of amino acid sequences:

- **IGORFs**
  We applied our fold state prediction method on the potential peptides encoded by the 105041 IGORFs extracted from the genome of *S. cerevisiae* with our program ORFtrack. In addition, we focused specifically on the 1235 occasionally translated IGORFs as well as on the 31 highly translated ones.

Therefore, we defined 2 additional datasets and applied the same protocol of size selection (20-70 amino acids) and redundancy elimination (30% identity) described for the dataset of the model construction.

- **Known folded peptides**
  157 peptides with experimentally characterized 3D structure were extracted from the class j of SCOPe database which contains fragments of longer proteins as well as short peptides (Chandonia et al. 2019).

- **Peptides encoded by AltORFs and bear experimental proof of expression**
  1935 peptides produced by the translation of ORFs in alternative reading frames of known human proteins (AltORFs) were extracted from the OpenProt database (https://openprot.org) (Brunet et al. 2019, 2021). This database offers a deeper view of the human proteome by annotating novel proteins. The AltORF-encoded peptides of our dataset cumulated translation evidence (through publicly available Ribo Seq datasets) as well as expression evidence (through publicly available MS-MS datasets) (Brunet et al. 2019, 2021).

5.2.3 Descriptors

Our prediction model was trained on a set of numerous sequence-based physicochemical descriptors calculated on the total amino acid sequence of every peptide (presented in Table 5.2). In more details, among the descriptors there were the frequency of each amino acid, the frequency of hydrophobic amino acids, the disorder and aggregation propensity of the sequence as well as the average value of the 58 AAindices described in the ProtFP descriptor set. These 58 descriptors are only based on the natural amino acids and have been selected so that they are largely independent by removing all the indices with large covariance (van Westen et al. 2013b, 2013a).

Table 5.2. Presentation of the various amino acid sequence-based descriptors calculated for every peptide sequence used for training our prediction model. The first column contains the name of the descriptor, the last column a more detailed explanation while the second column contains the type of the descriptor (**%** stands for frequency or sequence portion while **avg** stands for average value of the total sequence)

Name	Type	Explanation
X_frequency	%	Frequency of each one of the 20 amino acids (X stands for every amino acid)
Hydrophobic	%	Frequency of the strong hydrophobic amino acids (V,I,L,F,M,Y,W)
IuPRED	%	Portion of the sequence predicted as disorder prone (Explained in details in the methodology part of the manuscript)
IuPRED	avg	Mean IuPred value calculated for the total of the sequence
Anchor	%	Portion of the sequence predicted as disordered in solution but capable to fold upon binding
Aggregation	%	Portion of the sequence predicted as aggregation prone (Explained in details in the methodology part of the manuscript)
ARGP820103	avg	Membrane-buried preference parameters (Argos et al. 1982)
BHAR880101	avg	Average flexibility indices (Bhaskaran and Ponnuswamy 1988)
CHAM810101	avg	Steric parameter (Charton 1981)
CHAM820101	avg	Polarizability parameter (Charton and Charton 1982)
CHAM830101	avg	The Chou-Fasman parameter of the coil conformation (Charton and Charton 1983)
CHAM830107	avg	A parameter of charge transfer capability (Charton and Charton 1983)
CHAM830108	avg	A parameter of charge transfer donor capability (Charton and Charton 1983)
CHOP780201	avg	Normalized frequency of alpha-helix (Chou and Fasman 1978)
CHOP780202	avg	Normalized frequency of beta-sheet (Chou and Fasman 1978)
CHOP780203	avg	Normalized frequency of beta-turn (Chou and Fasman 1978)
CIDH920105	avg	Normalized average hydrophobicity scales (Cid et al. 1992)
FASG760101	avg	Molecular weight (Fasman 1975)
FAUJ880102	avg	Smoothed upsilon steric parameter (Fauchere et al. 1988)
FAUJ880103	avg	Normalized van der Waals volume (Fauchere et al. 1988)
FAUJ880104	avg	STERIMOL length of the side chain (Fauchere et al. 1988)
FAUJ880105	avg	STERIMOL minimum width of the side chain (Fauchere et al. 1988)
FAUJ880106	avg	STERIMOL maximum width of the side chain (Fauchere et al. 1988)
FAUJ880109	avg	Number of hydrogen bond donors (Fauchere et al. 1988)
FAUJ880110	avg	Number of full nonbonding orbitals (Fauchere et al. 1988)

FAUJ880111	avg	Positive charge (Fauchere et al. 1988)
FAUJ880112	avg	Negative charge (Fauchere et al. 1988)
FAUJ880113	avg	pK-a(RCOOH) (Fauchere et al. 1988)
GRAR740102	avg	Polarity (Grantham 1974)
JANJ780102	avg	Percentage of buried residues (Janin et al. 1978)
JANJ780103	avg	Percentage of exposed residues (Janin et al. 1978)
JOND920102	avg	Relative mutability (Jones et al. 1992)
JUNJ780101	avg	Sequence frequency (Jungck 1978)
KLEP840101	avg	Net charge (Klein et al. 1984)
KRIW790101	avg	Side chain interaction parameter (Krigbaum and Komoriya 1979)
KYTJ820101	avg	Hydropathy index (Kyte and Doolittle 1982)
LEVM760102	avg	Distance between C-alpha and centroid of side chain (Levitt 1976)
LEVM760103	avg	Side chain angle theta(AAR) (Levitt 1976)
LEVM760104	avg	Side chain torsion angle phi(AAAR) (Levitt 1976)
LEVM760105	avg	Radius of gyration of side chain (Levitt 1976)
LEVM760106	avg	van der Waals parameter R0 (Levitt 1976)
LEVM760107	avg	van der Waals parameter epsilon (Levitt 1976)
NISK800101	avg	8 A contact number (Nishikawa and Ooi 1980)
NISK860101	avg	14 A contact number (Nishikawa and Ooi 1986)
PONP800101	avg	Surrounding hydrophobicity in folded form (Ponnuswamy et al. 1980)
RACS770103	avg	Side chain orientational preference (Rackovsky and Scheraga 1977)
RADA880108	avg	Mean polarity (Radzicka and Wolfenden 1988)
ROSG850101	avg	Mean area buried on transfer (Rose et al. 1985)
ROSG850102	avg	Mean fractional area loss (Rose et al. 1985)
ROSM880102	avg	Side chain hydropathy, corrected for solvation (Roseman 1988)
WARP780101	avg	Average interactions per side chain atom (Warme and Morgan 1978)
WOLR810101	avg	Hydration potential (Wolfenden et al. 1981)
VINM940101	avg	Normalized flexibility parameters (B-values), average (Vihinen et al. 1994)
TAKK010101	avg	Side-chain contribution to protein stability (kJ/mol) (Takano and Yutani 2001)
MONM990201	avg	Averaged turn propensities in a transmembrane helix (Monné et al. 1999)
KOEP990101	avg	Alpha-helix propensity derived from designed sequences (Koehl and Levitt 1999)
KOEP990102	avg	Beta-sheet propensity derived from designed sequences (Koehl and Levitt 1999)
MITS020101	avg	Amphiphilicity index (Mitaku et al. 2002)
COSI940101	avg	Electron-ion interaction potential values (Cosic 1994)
PONP930101	avg	Hydrophobicity scales (Ponnuswamy 1993)
ZHOH040102	avg	The relative stability scale extracted from mutation experiments (Zhou and Zhou 2004)
ZHOH040103	avg	Buriability (Zhou and Zhou 2004)
BAEK050101	avg	Linker index (Bae et al. 2005)
CASG920101	avg	Hydrophobicity scale from native protein structures (Casari and Sippl 1992)

5.2.4 Training of the prediction model

For the model training step, 20 random sequences of every fold state were extracted for constructing a test set independent from the sequences used for the training of the model. These sequences are consequently only used for the model's performance estimation.

From the remaining sequences, for each fold state category we selected randomly up to 200 sequences. These selected sequences constituted our training set. Specifically, for the S3 and Small fold states, for which we had less than 200 sequences, we did not remove redundant sequences but rather weighted their respective contribution during the model training. Therefore, using the mmseqs tool (Hauser et al. 2016) we clustered the sequences of these sets according to their similarity with a threshold of 30% of sequence identity. The contribution of each sequence in the training of the model was then weighted based on its representation by homologous sequences. For example, if 4 sequences share more than 30% of sequence identity, each one gets a weighting value of 0.25.

The fold state prediction model was based on multiple two-class SVM classifiers using the Radial Basis Function (RBF) Kernel function. One SVM classifier was trained independently for each fold category in order to distinguish between sequences belonging to this specific fold category and sequences which do not. As a result, five distinct SVM classifiers were generated, each one aiming at predicting whether a given amino acid sequence corresponds or not to its specific fold category. The advantage of using 5 independent classifiers relies on the fact that a peptide can be unannotated if it does not correspond to any of the 5 categories. This enables us to identify peptides with unexpected fold states. The hyperparameters C and gamma for each SVM classifier were defined based on a grid search where multiple combinations of different C and gamma are tested and the combination with the best performance is finally selected (explained later). The hyperparameters used for every SVM classifier are presented in Table 5.3.

Table 5.3. Parameters used for every two-class SVM classifier.

	C	**gamma**	**Kernel**
IDPs	1	0.01	rbf
DIBS	13	0.003	rbf
Small	1015	0.0001	rbf
S3	1071	0.0006	rbf
TMs	1	0.001	rbf

Then the five independent predictions were combined in order to give one final prediction based on the following schema:

- **Unique prediction:** If a single predictor was positive while the four others were negative, then the prediction was the one of the positive predictor.
- **Multiple prediction:** If two or more predictors were positive, we preferred not to force any prediction but rather we created a "multiple" prediction class. We consider that sequences belonging to this multiple class might share similarities with different fold states and could be placed in the "twilight zone" of multiple fold categories.
- **No prediction:** If none of the predictors was positive, then the sequence is assigned as "non-predicted" and correspond to cases that do not have any clear representative fold category in our training dataset.

The prediction schema is presented in Figure 5.2.

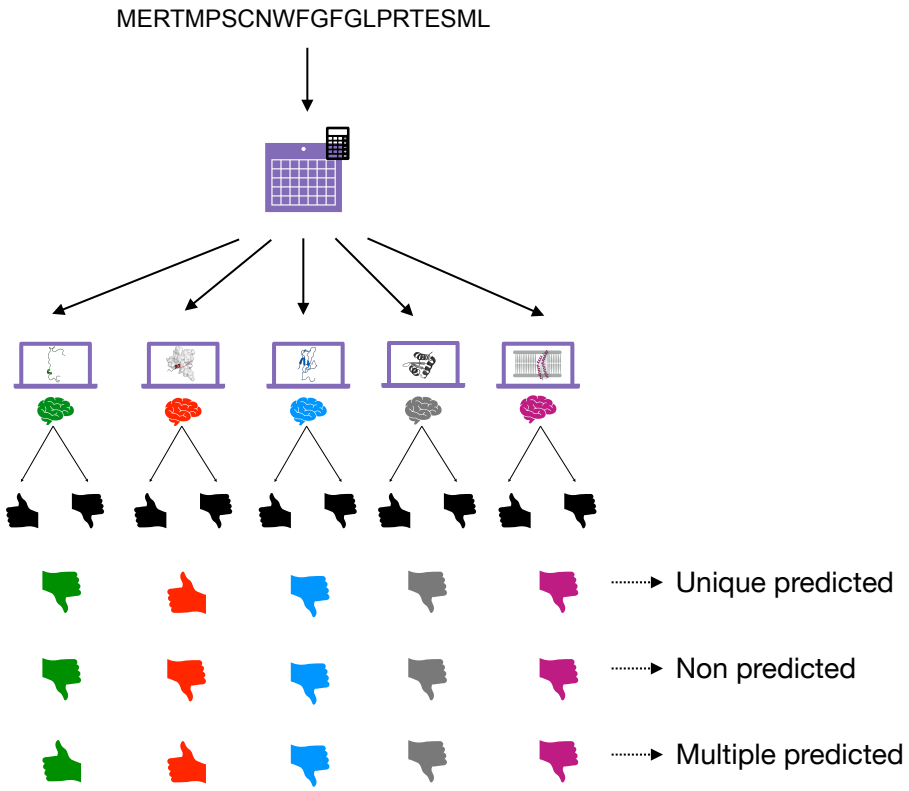


Figure 5.2. Representation of the two-class SVM-based prediction schema for the prediction of the fold state of a given short amino acid sequence. Starting with the amino acid sequence whose fold state is to be predicted, 83 physicochemical descriptors are calculated and are tested with the five SVM models each one trained to recognize one specific fold state (IDPs in green, DIBS in red, Small in blue, S3 in black and TMs in purple). Based on the prediction of every model our schema makes the final prediction (arrows on the right). In the first case, only the DIBS predictor was positive and consequently the final prediction will be uniquely DIBS. In the second case, none of the predictors made any prediction so the schema will not take any decision and will assign the sequence to the non-predicted class. In the third case, two predictors were positive (IDPs and DIBS) and so the sequence is assigned to the multiple class where more than one prediction was made.

5.2.5 Estimation of the model's performance

After having trained the five SVM classifiers, we tested their predictive performance by calculating different performance estimators. Therefore, a confusion matrix is calculated based on the output of the classifier and counts the number of its True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions.

Table 5.4. Example of a confusion matrix for a two-class classifier

		Prediction	
		Positive	Negative
Real case	Positive	Number of **TP**	Number of **FN**
	Negative	Number of **FP**	Number of **TN**

With this confusion matrix we can calculate for each classifier its precision, recall and F1-score estimators.

Precision is the fraction of the True Positive from the total instances predicted as positive. The precision is intuitively the ability of the classifier not to label as positive a case that is negative.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall is also termed sensitivity and is defined as the ratio of the True Positive to the number of actual positive cases. The recall is intuitively the ability of the classifier not to label as negative a case that is positive and thus to find all the positive cases.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F1-score is the harmonic mean of precision and recall and is considered one of the best metrics for classification models as it combines the information of these two metrics.

$$\text{F1} = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

These three metrics were used to estimate the performance of each SVM classifier (i) through a 5-fold Cross Validation (CV) process of 1000 iterations and (ii) on the independent test set. During one iteration of the CV, the training set is divided into five equal random parts. One part is excluded from the model's training process in order to serve as an independent test set for the performance estimation. The four remaining parts are used for the model's training and the performance estimators are calculated on the one part left aside. This process is repeated five times such that each part is used once as a test set. Consequently, at every iteration step of the CV five different models are generated each one combining four random parts of the initial training set. The average performance of all the five models is reported as the performance of every iteration step. This procedure is iterated 1000 times and the average performance of all the iterations is calculated as the final CV performance.

After combining the information from the five SVM classifiers we construct a larger confusion matrix which contains five classes instead of two (matrix 5x5). In order to assess the performance of the overall prediction model we calculate the same metrics (precision, recall and F1-score) for each class, and calculate their average, weighted by the number of true instances for each class.

5.3 Results

We generated a training set of 986 amino acid sequences of five different fold state categories. For every sequence, 83 sequence-based physicochemical descriptors were calculated in order to train our fold-category classification model (see Methods for details). At note, we used descriptors calculated strictly with the peptides' amino acid sequence and not with their 3D structure, so that the prediction model can be applied to every peptide sequence.

5.3.1 The physicochemical descriptors can discriminate the fold categories

First, we investigated whether the descriptors selected presented a discriminative capacity among the five categories of fold states. Non informative descriptors would generate models with low predictive accuracy. To do so, we performed a Principal Components Analysis (PCA) on our training set (Figure 5.3). The PCA is a dimensionality-reduction method which transforms a large set of variables into a smaller one that contains most of the information in the large set (Abdi and Williams 2010). Principal Components (PCs) are new variables that are constructed as linear combinations of the initial ones. These combinations are done in such a way that these new variables are uncorrelated and most of the information within the initial variables is compressed into the first PCs. The PCs correspond to novel axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible (Abdi and Williams 2010).

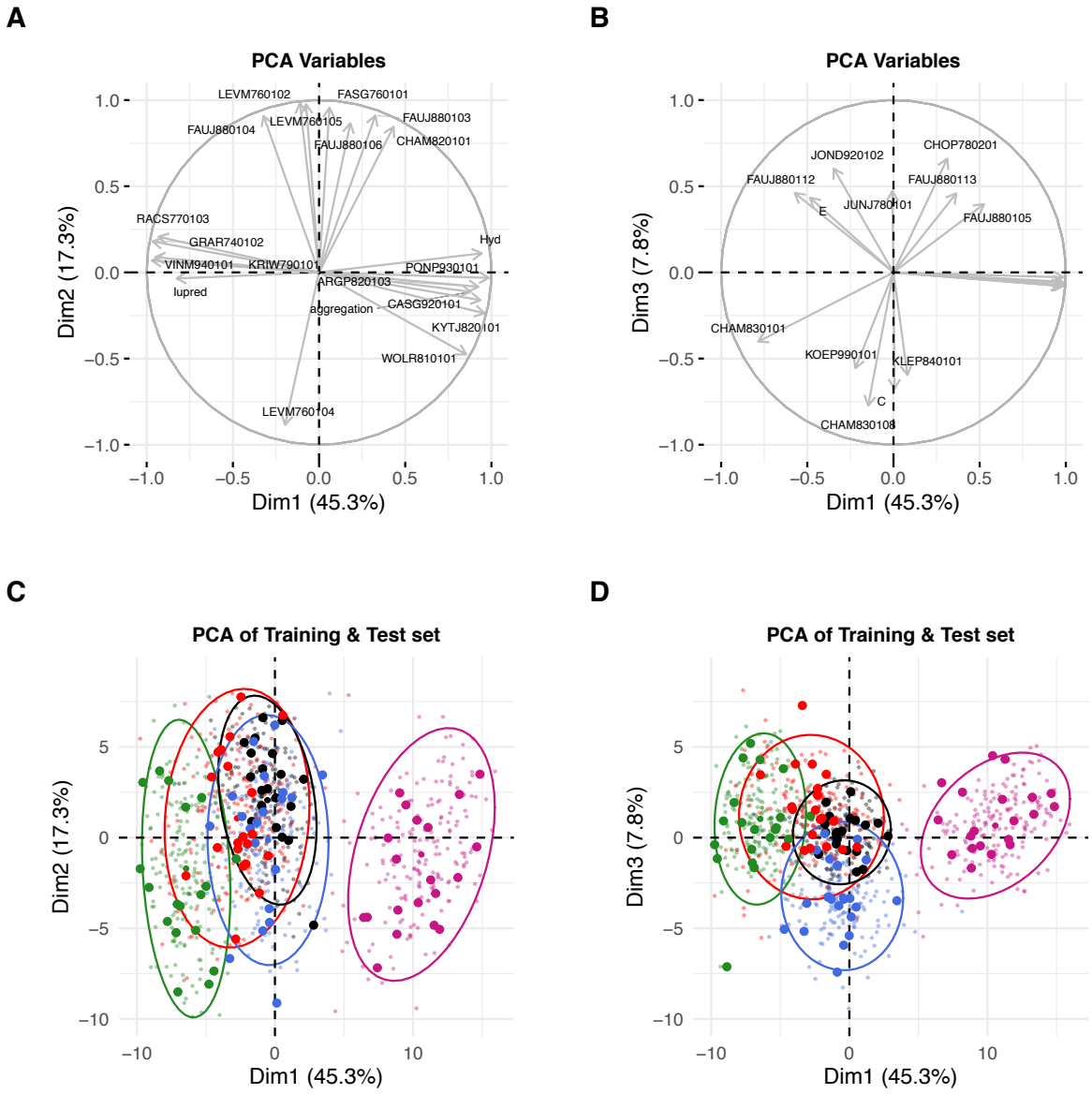


Figure 5.3. Principal Components Analysis of the training set with projection of the test set. (A) Correlation circle between the variables of the first and the second PCs. Only variables with high contribution are presented. **(B)** Correlation circle between the variables of the first and the third PCs. Only variables with high contribution are presented. **(C)** PCA of the first and the second PCs. The individuals of the training set are presented with ellipses and small points colored by their fold state category. The level of the ellipses is set to 90% of the data per category. The test is projected on the PCA with large-solid points and colored by fold state category. **(D)** PCA of the first and the third PCs. The individuals of the training set are presented with ellipses and small points colored by their fold state category. The level of the ellipses is set to 90% of the data per category. The test set is projected on the PCA with large-solid points and colored by fold state category. **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.**

In Figure 5.3 is represented the PCA of the three first PCs of the training set. Altogether the three PCs explain up to 70.4% of the overall data variance. One can observe that the two first PCs can clearly separate the two extreme fold categories, i.e., the IDPs (in green on the left part of the plot) and the TMs (in purple on the right part of the plot), reflecting clearly their opposite character according to the PC1 axis. However, the remaining three categories locate at the central part of the plot in-between the two other extreme fold categories (IDPs and TMs). According to PC1 and PC2, it is difficult to distinguish them as shown by their overlapping ellipses (red, blue and black for DIBS, Small and S3, respectively). This result could reflect a continuum among the different categories of fold states. Indeed, the PC1 consists mostly of descriptors which are associated with the hydrophobic content of the amino acid sequences. In more details, the negative values of the PC1 (distinguishing the IDPs) are associated with peptides presenting high polarity (GRAR7401102, RACS770103), high flexibility (VINM940101), high percentage of exposed residues (JANJ780103) and high disorder propensity. On the contrary, the positive values of the PC1 (distinguishing the TMs) are associated with peptides presenting high frequency of strong hydrophobic residues (V, F, I, L, W, M), high aggregation propensity, high preference to get buried in membrane (ARGP820103), high hydropathy index (KYTJ820101) or hydrophobicity scales (PONP930101, CASG920101) and high relative hydration potential (WOLR810101) corresponding to peptides with low solubility. To sum up, the PC1 axis separates amino acid sequences based on their potential for being disordered and rich in polar residues (adopting lower values), their propensity to aggregate and being insoluble (adopting higher values) or presenting intermediate levels of these properties, and thereby expected to be foldable. Finally, the PC1 axis is in line with the HCA foldability score which mostly rely on the distribution and patterns of hydrophobic and hydrophilic residues. Interestingly, Figure 5.3D shows that the PC3 axis provides additional information compared with HCA, since it enables the discrimination between the Small fold category and those of DIBS and S3, though the S3 category overlaps partially the one of DIBS. The study of the main descriptors of the PC3 revealed that the Small fold category presents high average value of net charge (index KLEP840101) and high frequency of cysteines. These results may explain the less ordered character of this fold category, due to its higher net charge content. However, the higher frequency of the disulfide bond-formatting amino acid cysteine could favor the formation of stabilizing interactions leading to a fold that is overall foldable. All these

observations described on the training set make us confident that the 83 selected physicochemical descriptors, are capable of capturing differences among the different fold states. An important descriptor is the hydropathy of the sequences which is able to separate the IDPs and the TMs from the three other fold categories. In addition, the net charge and the frequencies of amino acids seem to participate in better clarifying the groups. All the observations made on the training set were also valid for the test set presented in Figure 5.3C-D with large colored points.

An interesting observation based on the PCA plot, is the existence of a region between the ellipse of the TMs and the ones of the remaining fold states which presents very low density of points. This region has not been associated with any fold state but mostly corresponds to a transitory region between foldable in solution and highly hydrophobic aggregation-prone sequences. Knowing that the PC1 axis discriminates the sequences based on their hydropathy, sequences localized in this intermediate region are expected to present an important hydrophobic content but not as high as the TM domains. A very exciting question that arises based on these observations is whether this region free of specific fold category is simply a bias resulting from the choice of our five datasets of reference (i.e., the five categories may not cover all fold states observed in databases or in the nature) or whether it reflects a real gap in the structural space (i.e., functional peptides and/or peptides resulting from regulated or pervasive translation cannot populate this region).

5.3.2 Machine learning model performance estimation

Then we set-up a method which could either assign a given short amino acid sequence to one of the five fold states or to label it as non-predicted when the sequence could not be assigned to one of the five categories. To do so, we created a supervised prediction schema based on five independent two-class SVM classifiers, each one trained to predict one of the five predefined fold categories (see Methods for details). In Table 5.5 are presented the performance estimators calculated for the five two-class SVM classifiers on the independent test set and with cross validation on the training set. The five SVM classifiers present high performances, especially on the independent test set which has

not been used for the training procedure, enforcing the robustness of the prediction model.

Table 5.5. The performance estimators for the two-class SVM classifiers for the five fold categories with cross validation (CV) on the training set and on the independent test set.

	IDPs		DIBS		Small		S3		TMs	
	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
Precision	0.96	0.95	0.85	0.91	0.97	0.97	0.93	0.94	1.0	1.0
Recall	0.95	0.95	0.72	0.91	0.97	0.97	0.92	0.93	1.0	1.0
F1-score	0.95	0.95	0.74	0.91	0.97	0.97	0.92	0.93	1.0	1.0

Based on the five predictions of the SVM classifiers, our prediction schema is capable at assigning a given amino acid sequence either with one of the defined fold categories (IDPs, DIBS, Small, S3, TMs) or with "multiple" classes where more than one fold categories could be predicted. In addition, our prediction schema, contrary to other multiclass predictors, is also capable of not assigning any fold category if the prediction is not highly accurate. This option is very interesting as it prevents us from misassigning a peptide that does not belong to any of the five categories which is very important (i) for studying the IGORFs that may exhibit different fold states from peptides stored in structural databases and (ii) for considering the gap observed previously on the PCA, that may be populated by unevolved sequences.

Table 5.6 presents the performance estimators calculated for the prediction schema on the independent test set. It must be mentioned that the multiclass predictions, by definition are not associated with a single fold category (as they contain more than one category predictions) and as a result always count as a negative prediction, thereby overestimating our negative predictions. Overall, the fold state prediction model presents high performance on the independent test set which becomes even higher if we consider as positive the multiclass predictions whose real fold category made part of their predicted fold states, thus supporting the important capacity of our schema at correctly predicting the fold potential of short amino acid sequences.

Table 5.6. Performance estimators for our prediction schema on the test set before and after the multiclass prediction reannotation.

	Before multiclass reannotation	After multiclass reannotation
Precision	0.93	0.94
Recall	0.82	0.90
F1-score	0.87	0.92

The prediction occurrences on the test set per category are presented explicitly in Table 5.7. It can be observed that DIBS is the most difficult fold category to be predicted and indeed, most of the times it is confused with either IDP or S3 or assigned to multi-classes precisely involving S3 and DIBS fold states. This observation reflects probably the special behavior of the DIBS category presenting a transitional character between disorder and folded state depending on their environment (in solution or upon interaction with another protein, respectively). This is also supported by the PCA plot where the DIBS ellipse is placed between the IDPs and the S3 fold states presenting important overlap with both categories.

Table 5.7. The prediction results in absolute numbers of our fold state prediction model on the independent test set.

	True Positive	**False Positive**	**Multiclass Predicted**	**Not predicted**
IDPs	17	1	1	1
DIBS	11	3	4	2
Small	18	0	1	1
S3	16	2	2	0
TMs	20	0	0	0

5.3.3 Prediction on known folded peptides

The performance estimators of the prediction schema on the test set, highlight its capacity of predicting with high confidence the fold state of a short amino acid sequence. Even though independent from the training set, the test set consists of peptides

corresponding strictly to one of the five defined fold states. In this part, we are predicting the fold states of a real dataset of 157 folded peptides (extracted from the class j of the SCOPe database) which have not been selected according to the definition of the five fold categories and consequently we have no *a priori* about their specific fold state. Their predictions are presented in Table 5.8. It can be observed that the model predicts only 3.8% (6/157) of the sequences as IDPs, while it does not predict any fold category for 22.3% (35/157) of the cases. The remaining 73.9% of the sequences were assigned to one or multiple fold categories.

Table 5.8. Count and percentage of the predicted fold categories for the 157 peptides.

Prediction	Counts	Percentage (%)
IDPs	6	3.8
DIBS	34	21.7
S3	24	15.3
Small	20	12.7
TMs	25	15.9
Multiple	13	8.3
Non-Predicted	35	22.3

In Figure 5.4 is presented the PCA on the training set (in colored ellipses) together with the projection of the 157 peptide sequences colored according to the predictions of the model. With only exception the TMs, the different fold state predictions are well positioned inside or around their corresponding ellipse delimitations supporting once more that the predictions made by the model are quite accurate.

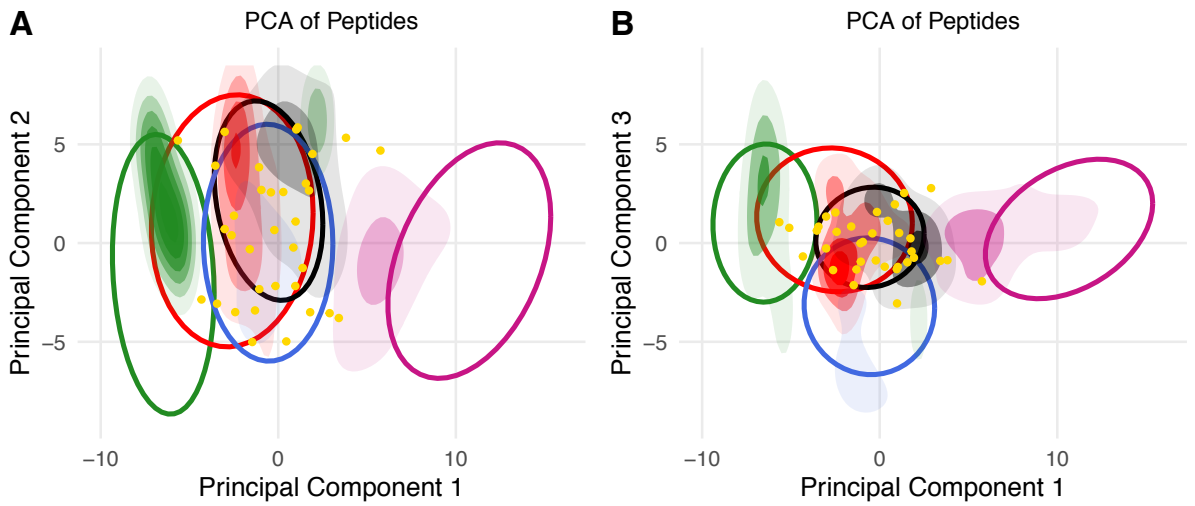


Figure 5.4. Principal Components Analysis of the training set presented with colored ellipses (ellipse level at 90% of the data) and projection of the 157 peptides extracted from the class j of SCOPe database (in colored densities). **(A)** PCA of the first and the second PCs; data colored according to the fold state prediction made by our model **(B)** PCA of the first and the third PCs; data colored according to the fold state prediction made by our model. The non-predicted cases are projected in yellow points. **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.**

At note, 60% (15/25) of the TM predicted cases are found to be delocalized from their corresponding ellipse and populating the region free of fold state. Research of these 15 peptides on the PDB revealed that all of them are indeed TM peptides whose structure has been characterized in apolar solvents resembling the membrane hydrophobic environment. This interesting result shows that the prediction model is quite accurate at predicting transmembrane peptides even though their descriptors combinations make them fall out of their corresponding ellipse defined by the training set. This suggests that our model has captured the TM propensity of these peptides even though they display different behavior (according to the first three PCs) from the TM peptides of the training set. On the other hand, the 20 Small fold state predicted cases are found to be well localized inside their corresponding ellipse (in blue) while the vast majority of their PDB structures correspond to cysteine-rich structures which stabilize mostly grace to disulfide bonds formation. The scorpion toxins androctonin and kappa-Hefutoxins as well as various conotoxins are some of the examples of cysteine-rich peptides that are predicted to participate in this fold category. Concerning the peptides predicted as the three other fold categories (IDPs, DIBS and S3), they are found well localized inside their

corresponding ellipses (green, red and black, respectively) even though some (7/24) of the peptides predicted as S3 fold state are outside the border of the black ellipse in the region free-of-fold-state.

The 3D structures and the PDB information for the peptides predicted as DIBS (34) and S3 (24) fold state, revealed that both fold categories contain mostly amphipathic peptides which can either interact with other receptor proteins (such as hormones) or embed their helical structure at the surface of membrane bilayers, burring in both cases their hydrophobic part. The similar character of these two fold categories is also supported by the important overlapping of their respective ellipses on the PCA plot. Indeed, the S3 fold state ellipse (in black) is included inside the ellipse of the DIBS fold state (in red). In order to better understand the predictions made by our model, we investigated the experimental conditions in which every peptide was characterized. We defined three types of structures according to the information in their corresponding PDB file:

(a) **Stable in solution:** Peptides with regular secondary structure(s), detected without any partner, in water solution or other polar solvent mimicking the cytoplasmic environment.
(b) **Stable upon interaction:** Peptides with regular secondary structure(s) detected either in interaction with another protein or in mixtures of polar-apolar solvents (i.e., micelles) mimicking the membrane surface environment. For these cases we have no information about their structural state alone in solution.
(c) **Unstable in solution and stable upon interaction:** Peptides which bare proof of their unfolded state in water solution and detected with regular secondary structure(s) in interaction-mimicking environments.

One should notice that the categories b and c are quite similar, however the cases in category b do not mention explicitly the disordered state of the peptide under aqueous solution.

Interestingly, 62.5% (15/24) of the S3-predicted cases are found as stable in solution in a monomeric form (category a) while the remaining 37.5% are found stable upon interaction (categories b and c - 33.3% and 4.2%, respectively). On the contrary, 35.3%

(12/34) of the DIBS-predicted peptides present regular secondary structure(s) in solution while 64.7% (22/34) are found stable upon interaction (categories b and c). At note, 44.1% (15/34) of the DIBS-predicted cases, bear proof of their unstructured state under solution (category c). Finally, most of the S3-predicted peptides were characterized in solution while most of the DIBS-predicted ones were characterized in conditions that impose fold stabilization (mimicking membrane surfaces). Interestingly, despite the fact that most of these peptides (predicted as S3 or DIBS) correspond to similar 3D structures (i.e., amphipathic helices), our model was able to capture the conditions necessary to acquire their 3D fold.

Finally, 35 peptides were not predicted with any fold category and their projections are presented with yellow points in Figure 5.4. They are localized at the central part of the plot, falling mostly in the S3/DIBS or Small fold state ellipses (88.6% - 31/35), while 11.4% (4/35) are localized in the free-of-fold-state region. Research on the PDB revealed that 31.4% (11/35) of the non-predicted cases are found stable in solution (category a), 45.7% (16/35) are stable upon interaction (category b) while 22.8% (8/35) stabilize through interaction while present an unstructured state in solution (category c). Their information from the PDB in combination with their central positioning on the PCA plot, support that the non-predicted cases correspond to an heterogenous population of peptides that resemble strongly to S3 and/or DIBS fold states. However, these cases may be characterized by descriptor values which are distinct from the ones of the peptides used for the training. This would explain the failure of the model to predict their fold category with accuracy.

These observations highlight two coexisting difficulties which probably cause our model to fail at taking a decision and pose an open question concerning the sensitivity of our prediction schema. The first one is the intrinsic similarity (in terms of structural characteristics) that share the S3 and DIBS fold states as discussed previously. However, we showed that although, most of the peptides predicted as S3 or DIBS correspond to amphipathic helices, our model was able to identify their respective folding properties (i.e., stable in solution in a monomeric form or folded upon binding with a partner or a membrane). Nevertheless, a small fraction of the peptides predicted as S3 or DIBS fold states were inverted between these two categories. This reflects a subtle continuum

between these two fold state categories whose borders according to our sequence-based descriptors overlap as shown with the PCA of the training set (Fig 5.3). The second difficulty lies in the lack of sensitivity of our model as shown by the 35 peptides that were annotated as "non-predicted" while our manual inspection revealed that they mostly corresponded to peptides folded in a monomeric form or able to fold upon binding with a partner. We can ask whether the peptides used for the training of the SVM models did not display enough diversity, thereby explaining that our models face difficulty at generalizing with peptides that display different distributions of features from those used for the training. Even though the 35 peptides that were classified as non-predicted mostly fall in the ellipses of the S3, DIBS or Small categories, thereby reflecting that they display similar descriptors according to the 3 first axes of the PCA, they may display different values for the remaining descriptors. This reflects an important limitation of our method which maybe is quite accurate at detecting peptides that present similar descriptor values to those of the peptides used for the training set but struggle to handle further variability. Enlarging the training set and monitoring the impact of the variability in the accuracy of the predictions is to be further investigated.

5.3.4 Fold prediction on the IGORFs

In this section, we launched our fold state prediction schema on the peptides encoded by the IGORFs of *S. cerevisiae* in order to study the distribution of the different fold states in the ensemble of the IGORFs. As a reference, we also calculated their HCA foldability score and compare them with the fold state predictions. In Table 5.9 are presented the results of the fold category predictions for the set of IGORFs, in total but also grouped by their HCA score. 28.4% (29873/105041) of the IGORFs were predicted as belonging to the TM fold category, 42.8% (44965/105041) were predicted with one (or multiple) fold states capable of acquiring a 3D fold in solution or upon interaction (DIBS, S3, Small or combination of them) while 0.8% (998/105041) were predicted as IDPs. In addition, 27.8% (29205/105041) of the IGORFs were not assigned to any of the fold states. These results reflect a large range of fold states that exist in the IGORFs of *S. cerevisiae* with an important depletion of intrinsically disordered peptides which has already been highlighted with the HCA foldability score.

Table 5.9. Frequency of the fold state predictions for the set of IGORFs in total but also grouped by their HCA score.

Prediction	Total (%)	Low HCA (%)	Intermediate HCA (%)	High HCA (%)
IDPs	0.8	10.0	0.1	0
DIBS	14.7	48.1	15.4	2.7
S3	10.6	3.3	14.0	4.2
Small	11.5	6.9	13.5	7.9
TMs	28.4	0.2	18.6	62.3
Multiple	6.0	12.0	5.4	6.0
Non-Predicted	27.8	19.5	33.0	16.7
Total	100	100	100	100

In Figure 5.5 is presented the PCA of the training set (in colored ellipses) together with the projection of the total IGORF-encoded peptides colored according to their fold state prediction.

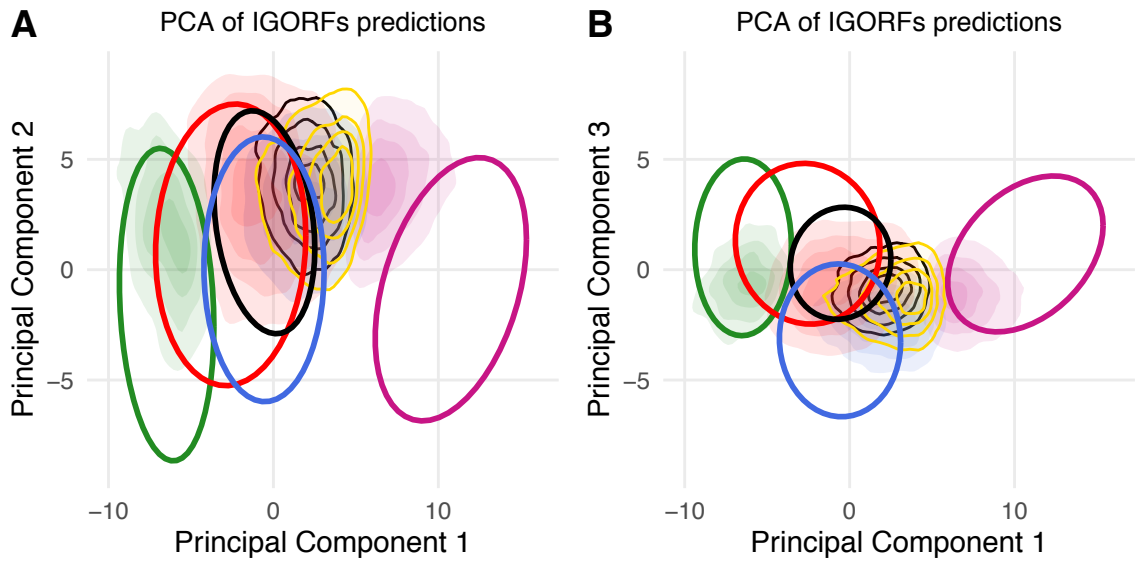


Figure 5.5. Principal Components Analysis of the training set presented with colored ellipses (ellipse level at 90% of the data) and projection of the total IGORFs (in colored densities). **(A)** PCA of the first and the second PCs; data colored according to the fold state prediction made by our model **(B)** PCA of the first and the third PCs; data colored according to the fold state prediction made by our model. The non-predicted

cases are projected in yellow densities. **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.**

In Figure 5.5 we can see that the ensemble of the IGORFs presents a large dispersion all over the PCA plot highlighting the important fold potential diversity existing in the peptides encoded by the noncoding genome. Again, we can observe the continuum of fold states all along the PC1 axis which reflects the general hydrophobic content of the amino acid sequences. IGORFs predicted as IDPs and DIBS are found well localized inside their respective ellipses on the PCA plot. On the contrary, IGORFs predicted as TM, S3 or Small are found delocalized from their respective ellipses (ellipse level set at 90% of the points) with a tendency to cumulate towards the free of fold state region. Precisely, 47.9% (5349/11166) of the S3-predicted IGORFs and 45.4% (5512/12146) of the Small-predicted ones are found inside the region free-of-fold-state while it is the case for up to 71.4% (21336/29873) of the TM predicted IGORFs. Furthermore, almost one fourth of the total IGORFs (29205/105041) were not assigned to any fold category with 60% of them (17667/29205) localized at the fold state-free region.

In line with our previous work, these results reflect the important fold state variability existing in the peptides encoded by the yeast IGORFs, which ranges from completely disordered peptides until highly hydrophobic TM ones, prone to aggregate in solution. Even though the prediction model presents good predictive capacity on the test set and the independent set of 157 structured peptides, the PCA plot reveals that many IGORF-encoded peptides are delocalized from the fold state ellipse they were predicted to belong, towards the region free-of-fold-state. In fact, we showed with the 157 peptides dataset, that an important fraction (60%) of peptides experimentally shown to be TM, were localized outside the TM purple ellipse on the PCA, reflecting that our model was able to capture their TM propensity even though they display slightly different values for the descriptors associated with the 3 first PCA axes from those of the peptides of the training set. However, we do not know whether this holds for the other fold state categories. In particular, one can ask whether the peptides predicted as S3 or Small which are localized outside the S3 and Small ellipses respectively are predicted correctly or constitute false positives. More generally, one can ask whether our model trained on real

peptides is able to predict the fold state of unevolved peptides. Indeed, our model was trained on peptides' sequences which correspond to evolutionary optimized peptides that must satisfy the structure-stability-function compromise while the IGORFs, correspond to unevolved sequences. In other words, coding peptides have evolved to maintain their function including their interactions with their partners and more generally with the cellular environment. Indeed, it has been shown that the crowded cell has shaped the surface interaction properties of proteins to prevent them to be trapped in non-functional interactions (Levy et al. 2012; Macossay-Castillo et al. 2019). These different constraints exerted on coding peptides are expected to leave "footprints" on their amino acids composition that may be reflected by the ellipses. On the contrary, IGORFs may be able to fold, but have not been optimized for a function including interaction with specific partner(s) but also with the cellular environment. This may explain the fact that even though our method has predicted them as S3 or Small fold states, these unevolved peptides whose interaction properties with the cellular environment have not been optimized, fall outside the ellipses of the corresponding fold state categories. As a result, applying a prediction model trained to recognize fold states of coding sequences on a dataset of unevolved noncoding sequences is a very difficult task and thus, we cannot estimate to which extend the predictions made on the IGORFs are indeed highly accurate.

5.3.5 Comparison of the IGORF predictions with the HCA foldability score

In Figure 5.6A is presented the distribution of the fold states prediction on the IGORFs in the three distinct HCA score bins (Low, Intermediate and High). It can be observed that, as expected, sequences with low HCA scores are mostly predicted as IDPs (10%) and DIBS (48.1%) (one proportion z-tests p-values with intermediate and high HCA score bins: $< 2 \times 10^{-16}$) while sequences with high HCA scores are mostly predicted as TMs (62.3%) (one proportion z-tests p-values with low and intermediate HCA score bins: $< 2 \times 10^{-16}$). IGORFs with intermediate HCA scores are predicted with a large variability of fold states (from DIBS to TMs) and notably are enriched in S3-predictions (14%) (one proportion z-tests p-values with low and high HCA score bins: 1×10^{-4} and 6×10^{-4}, respectively), Small predictions (13.5%) (one proportion z-tests p-values with low and high HCA score bins:

1×10^{-2} and 4×10^{-2}, respectively) and also non-predicted cases (33%) (one proportion z-tests p-values with low and high HCA score bins: 1×10^{-3} and 6×10^{-4}, respectively).

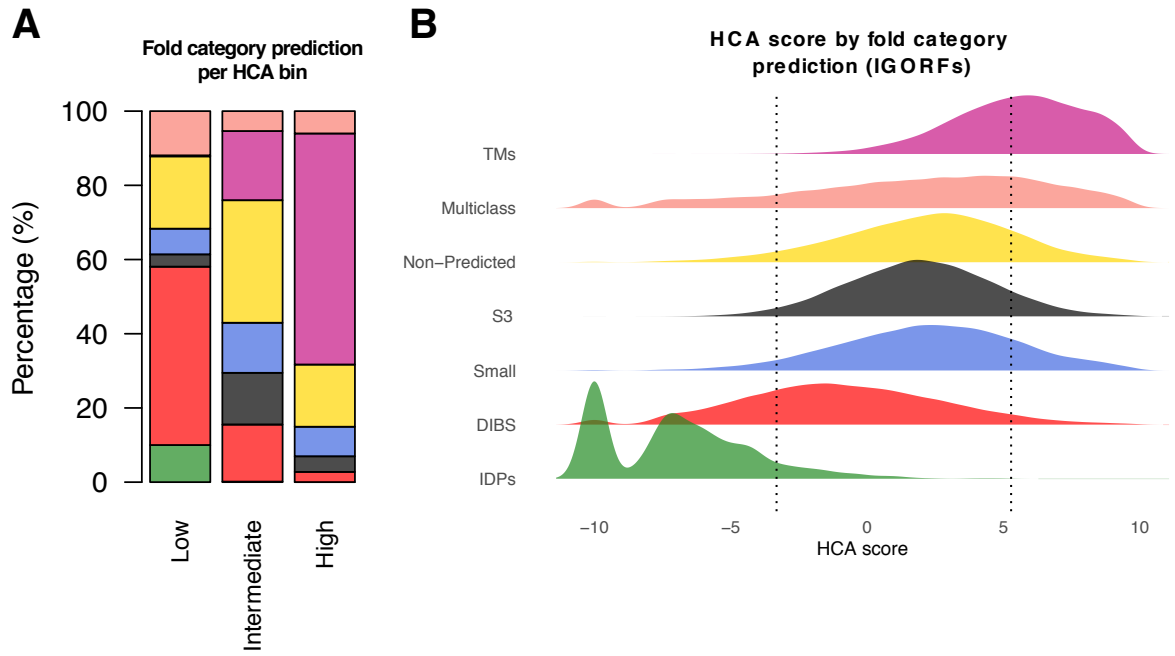


Figure 5.6. **(A)** Frequencies of fold state predictions for the IGORFs per HCA score bin. **(B)** HCA score distribution for the different IGORFs fold state predictions. Dotted black lines delineate the boundaries of the low, intermediate and high HCA score categories reflecting the three categories of fold potential (i.e., disorder prone, foldable, or aggregation prone in solution); **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3, Purple for TMs, Salmon for Multiple class prediction and Yellow for Non-predicted.**

Notably, in Figure 5.6B, is presented the HCA score distribution of the IGORFs grouped by their fold state prediction. The distributions of the HCA scores for the different fold state predictions on the IGORFs, present a continuum tendency which reflects the hydrophobic content of their amino acid sequences and consequently, their fold potential. Notably, IDP-predicted IGORFs present low HCA scores, DIBS-predicted IGORFs present low and intermediate HCA scores, S3 and Small-predicted IGORFs present intermediate HCA score while TM-predicted IGORFs present mostly high HCA scores. The multiple predicted cases present a wide range of HCA scores while the non-predicted cases present mostly intermediate scores supporting their potential ability to fold, though we do not know their fold properties (able to fold upon binding, stable in

solution or rudimentary fold). As a matter of fact, the HCA score distribution of the non-predicted cases was statistically similar to both S3 and Small predicted cases (two sided Kolmogorov–Smirnov test: $P = 7 \times 10^{-2}$ and $P = 4 \times 10^{-1}$, respectively). All the distributions present an important overlapping supporting that the limits between the different fold states are not distinct, highlighting a continuum among them. Overall, these observations reflect the important capacity of the HCA method at estimating the foldability potential of a given amino acid sequence. However, they also highlight that peptides with similar HCA scores can acquire different fold states and consequently support that our prediction model could be useful in order to further refine the results of HCA.

5.3.6 Translated IGORFs and human alternative ORFs present similar foldability potential

All the previous observations were made on the ensemble of peptides potentially encoded by the noncoding genome of *S. cerevisiae*. However, not all IGORFs are expected to be translated and to produce peptides in the cell. Among IGORFs that are translated, we showed that most of them are only translated occasionally. They are not expected to be functional, and we hypothesize that they will be short-lived in the evolutionary history. On the other hand, we showed that a small fraction of IGORFs, are translated with a strong signal. In particular, most of the reads (> 80%) that map on their genomic locus are in-frame, reflecting that the translation of these ORFs is strongly favored compared to the overlapping ORFs. This may indicate the optimization of their translation and finally the potential emergence of function. As a complement, we also investigated the fold state of peptides encoded by alternative ORFs overlapping with human CDS in different frames, observed with ribosome profiling and mass spectrometry experiments. These peptides, beyond the fact of being translated, are stable enough (i.e., not degraded) to be observed with mass spectrometry. They were observed in the human cells, and consequently are not directly comparable with the IGORFs translated in *S. cerevisiae* but they offer an opportunity to study real peptides, probably unevolved since they result from noncoding ORFs, and which are stable enough to be captured with mass spectrometry. They offer a great opportunity to interrogate whether unevolved peptides resulting from pervasive translation and stable enough to be observed with mass spectrometry, could populate the region free-of-fold-state which is populated by IGORFs but not by the peptides annotated

in the SCOP database. In Table 5.10 is presented the repartition of the fold state predictions made for every dataset.

Table 5.10. Fold state predictions made by our models on different datasets of peptides.

Prediction	IGORFs (%)	Occasionally translated (%)	Highly translated (%)	Human AltORFs (%)
IDPs	0.8	0.7	0	6.7
DIBS	14.7	14.9	29.0	16.6
S3	10.6	13.5	25.8	24.3
Small	11.5	9.8	9.7	9.6
TMs	28.4	24.1	9.7	5.6
Multiple	6.0	5.5	6.4	3.7
Non-Predicted	27.8	31.5	19.3	33.5

5.3.6.1 Occasionally and highly translated IGORFs fold state predictions

In Figure 5.7A-B is presented the PCA plot of the occasionally translated peptides colored based on their fold state prediction. The localization of the different fold state predictions on the PCA plot together with their repartition (Table 5.10), reveal that the occasionally translated IGORFs present a wide range of fold states which is similar to the one of the total IGORF peptides (one proportion z-tests, all p-values > 5×10^{-2}). Notably, similarly to the IGORFs, 31.5% (389/1235) of the occasionally translated peptides were not predicted with a fold state while 54.2% (211/389) of them were localized in the region free-of-fold-state. Also 43.7% (73/167), 46.3% (56/121) and 79.5% (237/298) of the occasionally translated IGORFs predicted as S3, Small and TM fold states, respectively localize outside their corresponding ellipses in the free-of-fold-state region. Again, these cases may reflect the fact that these unevolved peptides while belonging to these fold state categories display different values for the 3 first PCA axis descriptors from the peptides of the training set. All these results support that the occasionally translated IGORFs are IGORF-like without any fold state specificity. This suggests that they were not specifically selected to be translated according to their structural properties but may reach the translational machinery grace to their favorable genomic position or other reasons.

On the other hand, in Figure 5.7C-D are presented the PCA plots of the peptides encoded by the 31 highly translated IGORFs. The predictions of our model (Table 5.10) show that 70.9% (22/31) of the highly translated peptides are predicted with one (or multiple) fold states other than IDP or TM supporting that most highly translated IGORFs are able to fold in solution. In particular, highly translated IGORFs are clearly enriched in DIBS and S3 fold state predictions (one proportion z-tests p-values: 3×10^{-4} and 5×10^{-5} respectively) and depleted in TM predictions (one proportion z-tests p-values: 3×10^{-6}) compared to the predictions of the IGORFs. Moreover, 19.3% (6/31) of the peptides encoded by highly translated IGORFs do not present a fold state prediction with 50% of them (3/6) localized in the free-of-fold-state region. It is interesting to note that this fraction is less important than the one observed for the whole population of IGORFs suggesting that peptides belonging to the free-of-fold-state region are depleted in highly translated peptides. This region is expected to correspond to peptides with an important fraction of hydrophobic residues that may be deleterious in solution. This may explain the fact that peptides belonging to the free-of-fold-state region are under-represented in highly translated peptides. Could the fact that these highly translated peptides are enriched in peptides able to fold in solution or upon interaction be related with a potential functional character and consequently with the establishment of their more regulated expression? This is an interesting question which will need further studies.

To sum up, our results show that the occasionally translated IGORFs are IGORF-like presenting similar fold state predictions with the ensemble of IGORFs. On the contrary, the highly translated IGORFs are enriched in S3 and DIBS predictions and more generally enriched in sequences able to fold in solution or upon interaction with a partner, being mostly located in the central region of the PCA plot. This supports that IGORFs with important translation signal tend to display specific structural properties compared to IGORFs. Whether these properties were accompanied with the emergence of functionality and were a selected criterion is to be further investigated.

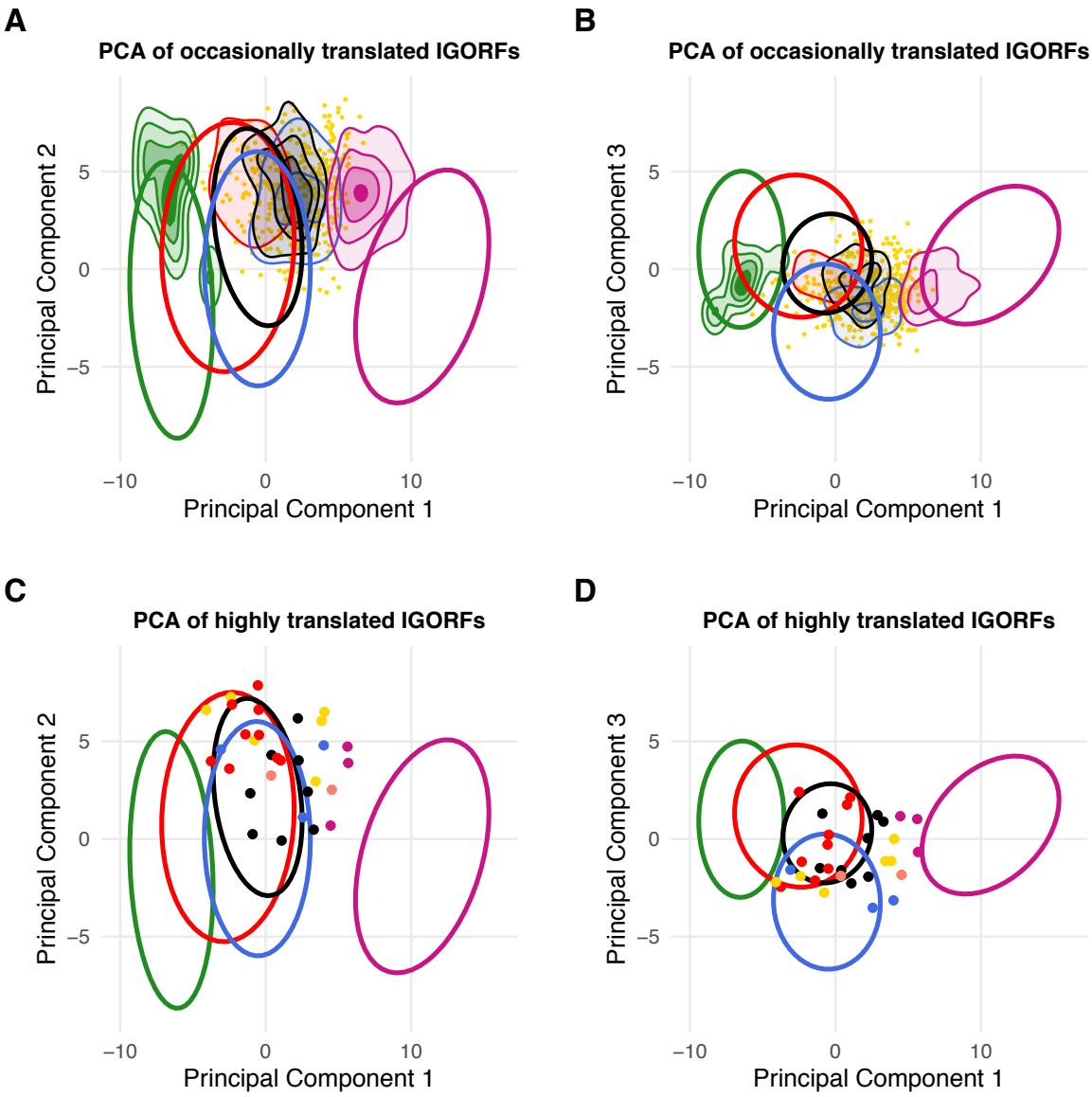


Figure 5.7. Principal Components Analysis of the training set presented with colored ellipses (ellipse level at 90% of the data) and projection of the 1235 occasionally and 31 highly translated IGORFs (in colored densities and points, respectively). **(A)** PCA of the first and the second PCs for the 1235 occasionally translated IGORFs **(B)** PCA of the first and the third PCs for the 1235 occasionally translated IGORFs **(C)** PCA of the first and the second PCs for the 31 highly translated IGORFs **(D)** PCA of the first and the third PCs for the 31 highly translated IGORFs; The non-predicted cases are projected in yellow points. All the data are colored according to the fold state prediction made by our model; **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.**

5.3.6.2 Human AltORFs fold state predictions

We showed that highly translated yeast IGORFs are not IGORF-like, being enriched in sequences coding potentially for peptides with DIBS or S3 fold state. However, being

translated does not necessarily involve being present in the cell. We can hypothesize that many of these products will be degraded right away. Therefore, this prompt us to take advantage of mass spectrometry data and investigate the fold state of peptides that were indeed observed experimentally. OpenProt is a database that contains numerous peptides annotated from the human genome (Brunet et al. 2019, 2021). These peptides are of great interest because they bear proof of expression with ribosome profiling and MS experiments and therefore correspond to peptides known to exist in the cellular environment while they may not be functional (i.e., resulting from pervasive translation). Using OpenProt, we extracted 1935 peptides produced by the translation of ORFs annotated in alternative reading frames of known proteins (AltORFs). In addition, one should note that their analysis cannot be directly compared with the one performed on the IGORFs of *S. cerevisiae* since they belong to different species and AltORFs, by definition, are not intergenic but overlap coding genes. They are thus expected to be subjected to different constraints and consequently may display different structure and sequence properties.

The predictions made by our model (presented in Table 5.10) show that 54.2% of the peptides are predicted with one (or multiple) fold states other than IDP (6.7%) or TM (5.6%) while 33.5% are not assigned with any fold state. In Figure 5.8 is presented the PCA plot of the AltORF-encoded peptides colored according to their fold state prediction. We can observe that the IDP- and the DIBS-predicted peptides are well localized inside their corresponding ellipses while on the contrary, TM predictions are clearly outside the purple ellipse as already observed with the dataset of 157 folded peptides. In addition, 28.9% (136/471) of the S3-predicted and 23.7% (44/186) of the Small-predicted peptides were positioned in the region free-of-fold-state although close to the border of their ellipses' delimitations. These results pose again the question concerning the accuracy of these predictions outside the ellipses and the overall specificity of the model on unevolved sequences.

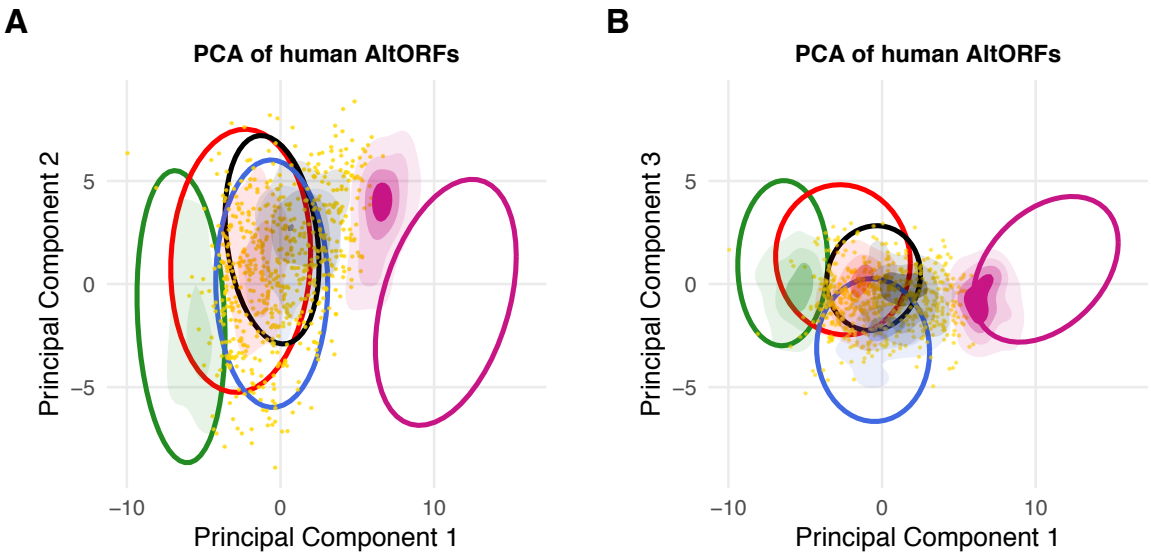


Figure 5.8. Principal Components Analysis of the training set presented with colored ellipses (ellipse level at 90% of the data) and projection of the 1935 AltORFs (in colored densities). **(A)** PCA of the first and the second PCs; data colored according to the fold state prediction made by our model **(B)** PCA of the first and the third PCs; data colored according to the fold state prediction made by our model; The non-predicted cases are projected in yellow points. All the data are colored based on the fold state prediction made by our model; **Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.**

Concerning the AltORF-peptides for which no prediction was made, they correspond to 33.5% (648/1935) of the total cases. Even though not directly comparable, one should note that this fraction is similar to the 27.8% of the non-predicted IGORFs (one proportion z-test, $P = 2 \times 10^{-1}$). However, it is interesting to observe that the non-predicted cases of AltORF-peptides are mostly positioned inside the fold state ellipses delimitations with only 22.2% (144/648) of them localized in the region free-of-fold-state (the respective number for the non-predicted IGORFs was 60%). This suggests again that AltORF-peptides observed with MS experiments are depleted in peptides belonging to the free-of-fold-state region, though it should be further investigated since these numbers are not directly comparable. On the other hand, it is unclear why an important fraction of peptides was not predicted with any fold state while located in the red, blue or black ellipses which reflect that they display similar properties according to those of the three first axes of the PCA. However, their positioning on the PCA plot permits us to hypothesize that an important fraction of them could share similarities with the

DIBS, the S3 or the Small fold states and therefore acquire a fold in solution or upon interaction with a partner.

In this section, we studied the predictions of fold state on a dataset of human peptides encoded by ORFs in alternative frames from protein coding sequences. These peptides have been proven experimentally to be produced in the human cells, but their functional role is not clear yet. Their overall fold state predictions together with their positioning on the PCA plot support that an important fraction of the AltORFs-encoded peptides may be able to fold in solution or upon interaction with a partner, while a minority of them are prone to aggregate. Although not directly comparable, these results recall the ones of the highly translated IGORFs of *S. cerevisiae* and consequently it is interesting to speculate that peptides resulting from IGORFs with higher level expression, or which are indeed observed in the cell display a high propensity for being folded in solution or upon interaction with a partner.

5.4 Conclusions

In this chapter we aimed at developing a method for the prediction of the fold states of peptides (20 – 70 amino acids) in order to further characterize the foldability potential of peptides encoded by IGORFs. One should notice that we are interested in predicting the foldability properties of a peptide, which can further provide us with information about its potential behavior in the cell and its global properties rather than having a precise description of its 3D structure. To do so, we constructed a dataset of peptides' sequences corresponding to five different fold state categories which present an important variability of fold potential, ranging from completely disordered peptides in solution until transmembrane peptides prone to aggregate in solution but probably capable to fold in a lipidic bilayer. The fold state prediction model performed quite well on the independent test set presenting F1-score of 0.87. In addition, when applied on a dataset of known structured peptides, our model predicted one or multiple fold states for 77.7% of the cases. Manual inspection of the structural content and the experimental conditions in which these peptides were characterized, revealed that our model was correctly predicting all the TM cases and the majority of DIBS and S3 cases. However, it

is not clear why our model did not predict any fold state for 22.3% of the peptides, as our manual inspection supports that they mostly resemble DIBS or S3 peptides. As a result, further studies on the descriptors' distributions are needed, in order to better understand why our model failed to give an accurate fold state prediction for these cases.

In order to better understand the predictions of the model, we studied the capacity of our descriptors to discriminate the different fold states by performing a Principal Components Analysis on the training set. Our results reveal that the different fold categories can be separated according to the first three PCs. However, DIBS, S3 and Small fold states overlap partially reflecting a continuum of properties between these three foldable categories (to note that they all correspond to intermediate HCA scores). This continuum reflects the general hydrophobic content of the peptides and consequently their fold potential. On the contrary, disordered peptides (with low hydrophobic content) as well as TMs (with high hydrophobic content) are well separated from the other categories. The remaining three categories, with intermediate hydrophobic content, locate in-between. Interestingly, the DIBS presented a large dispersion which overlapped with the IDPs and the S3 fold states, showing that the DIBS category can host a large variability of peptides with different levels of fold potential. In addition, peptides of the Small fold category can be discriminated from those of the S3 category grace to their higher average net charge and enrichment in cysteines which may altogether participate in the fold properties specific to this category. For instance, knottins and toxins belong to this fold category. These peptides are known to be depleted in regular secondary structures and enriched in cysteines which may stabilize the overall structure of the peptide. The PCA plot constituted a very useful tool which helped us to have a global overview of the data, and which can be useful to *a posteriori* further analysis of the predictions.

Interestingly, the PCA plot revealed a region between the ellipse of the TM and the ones of the other fold states which was not occupied by any of the five defined fold states. This observation made us wonder whether this less occupied region on the PCA plot, corresponds simply to a technical bias due to the lack of an actual fold category from our initial training set, or is a real biological "gap" corresponding to an underrepresented subset of amino acid sequences in the natural proteomes. We can hypothesize that the

peptides falling in this free of fold state region, which display a high hydrophobic content (though lower than the TMs), could constitute the mirroring category of the DIBS. Indeed, globular proteins are characterized by a subtle equilibrium between hydrophobic and hydrophilic residues (Bresler and Talmud 1944). DIBS are characterized by a displacement of the equilibrium towards more hydrophilic residues and are thereby disordered in solution but able to fold upon binding. Peptides corresponding to the free-of-fold-state region are on the contrary enriched in hydrophobic amino acids. Whether sequences falling in this "empty" region correspond mostly to hydrophobic peptides, unstable in polar environments and which may stabilize only under oligomerization, embedded in a protein or in membrane surfaces, is an interesting question. Unlike DIBS, these hydrophobic peptides unable to fold as monomers, may become deleterious in aqueous environment such as the cytosol, explaining why they are underrepresented in the coding world.

Prediction on the IGORFs

Our model's predictions on the potential peptides encoded by the *S. cerevisiae* IGORFs, reveals a large range of different fold states harboring in the yeast noncoding genome. This is in line with the vast foldability potentials observed with HCA in section 3 of this manuscript. As a matter of fact, 72.2% (75836/105041) of the peptides are assigned with one (or more) fold states with strikingly 28.4% of the cases being TMs. Notably, IGORFs predicted as TM, S3 or Small fold states present an important delocalization from their corresponding ellipses posing questions about the accuracy of the model's predictions. For the rest 27.8% (29205/105041) of the IGORFs, our model could not make any prediction with 60% (17667/29205) of these non-predicted cases being localized in the free-of-fold-state region and the remaining 40% (11538/29205) being inside the ellipses of the different fold states. As discussed previously, whether this region free-of-fold-state correspond to a biological gap highlighting another population of potential peptides that are "forbidden" or "limited" in the coding world is an interesting question to which we do not have the answer yet. On the other hand, the other 40% of the non-predicted cases supports that, eventually an important portion of the non-predicted peptides could acquire a 3D structure, but it is unknown why our model is not capable at assigning them to a specific fold category. However, it must be noted that our prediction model was

trained on a dataset of fold categories with known structured and functional peptides which have long been subjected to evolutionary selection and thus have optimized their fold potential. Consequently, we are wondering to what extend we can extrapolate and use a model trained to detect the fold state of coding and evolutionary restrained peptides on a set of noncoding and unevolved sequences. This is an extremely difficult question to which we probably do not have the answer.

Expressed noncoding peptides present important fold potential

Predicting the fold states of the total set of IGORFs permits us to have an overall idea about the potential small folds that could be produced upon pervasive transcription and non-canonical translation. However, only a subset of the total IGORFs would be translated and finally produced as peptides in the cell. Having identified 1235 occasionally translated yeast IGORFs with weak translation signals, we observed that the reparation of their fold state predictions together with the dispersion of their points on the PCA plot were statistically non distinguishable from the ones of the total IGORFs. These observations are in line with the results presented in section 3, showing that the occasionally translated IGORFs present similar structure and sequence properties to IGORFs. These results permit us to speculate that IGORFs, no matter their structural properties or fold potential, can be translated in the cell and that their translation is independent from these features.

On the contrary, the 31 highly translated yeast IGORFs are enriched in S3 and DIBS predictions and their dispersion on the PCA plot is clearly restrained to the central region, contrarily to the occasionally translated IGORFs. The same tendency (even though more dispersed) is observed for 1935 human peptides produced by the translation of alternative reading frames of known protein ORFs (AltORFs). It must be mentioned that Bartonek et al. (2020) have reported the interdependence of the hydrophobicity profiles of protein sequences with the two other overlapping reading frames. Whether the important foldability potential of human AltORFs is related to a potential functional role in the cell, or this foldability is simply an intrinsic property inherited by their overlap with protein coding ORFs is an interesting question for which we do not have the answer yet. We can hypothesize that this strong translation signal indicates a functional outcome for these peptides. Whether the structural properties observed for the highly translated

IGORFs favor the emergence of function and the optimization of their translational activity is to be further investigated.

Functionality

Our results revealed that almost one third of the IGORFs were predicted as participating in the TM fold category believed to aggregate in solution but to be foldable in lipidic environments. Notably, proteins encoded by smORFs have been shown to be localized at the membranes of the cells presenting a broad range of functions (Orr et al. 2020). Several small proteins (in eukaryotic and prokaryotic organisms), presenting a wide sequence diversity, have been found to participate into larger membrane protein complexes such as the photosystems I and II or the cytochrome oxidases. In addition, small proteins have been found to act as positive or negative regulators of membrane proteins thus participating into the cellular responses to environmental changes (Orr et al. 2020). Another example of functionality was presented by Knopp et al. (2019) who identified three peptides encoded by randomly generated ORFs that increased significantly aminoglycoside resistance of bacterial cells. Combining genetic and functional analyses they showed that these highly hydrophobic peptides, once inserted into the membranes, reduce the membrane potential and as a result decrease the aminoglycoside uptake of the cell. This study constitutes a very interesting example of random DNA sequences (reminding the IGORFs) which are capable at encoding peptides that confer selective benefits to the organism and illustrates how expression of random sequences could spuriously lead to the origination of new genes. A very interesting question is how these highly hydrophobic peptides traverse the cytoplasm towards the cellular membranes without aggregating or getting stuck into non-specific interactions with other cellular proteins. Recently, Tretyachenko et al. (2021) studied a large random-sequences library and showed that many random peptides arrive to form interactions with the DnaK chaperone a mechanism which could potentially serve for the “safe” translocation of highly hydrophobic peptides towards the membranes.

Nevertheless, determining the function of peptides encoded by IGORFs is a very difficult and ambitious task which is out of the scope of this manuscript. In fact, we cannot know what a pervasively translated peptide really does in the cell but studying peptides with

known functions, we could investigate what their structural properties would permit them to do. In a preliminary analysis we defined three datasets of functionally annotated peptides from the SATPdb (Structurally annotated therapeutic peptides database) (Singh et al. 2016). They constitute of 1276 antimicrobial peptides, 662 toxic peptides and 120 peptides participating into the cell-cell communication. Our model revealed that the cell-cell communication peptides were clearly depleted in TM peptides (1.6%) while none of their 23 non-predicted cases was found located in the free-of-fold-state region (Table S5.1 and Supplemental Fig. S5.1). This highlights that IGORFs falling in the free-of-fold-state region or assigned as TMs do not share structural properties with cell-cell communication peptides and we can hypothesize that they will not give rise to such peptides. It is interesting to observe that the function of the cell-cell communication peptides is somehow not compatible with a high hydrophobic content. Indeed, this category of functional peptides consists of hormone peptides and quorum sensing peptides (Singh et al. 2016; Verbeke et al. 2017) which both activate their receptor protein through highly specific binding. We can hypothesize that these peptides present an increased hydrophilic content in order to avoid non-specific interactions and to ensure the specific interactions with their receptor. On the other hand, 52.1% (345/662) of the toxic peptides were assigned to the Small fold state category, while only 3% (20/662) of the cases are non-predicted and locate in the free-of-fold-state region (Table S5.1 and Supplemental Fig. S5.1). This highlights that IGORFs assigned to the Small fold category share similar structural properties with known toxic peptides and this could be a potential fate of these IGORFs. Finally, the antimicrobial peptides revealed a wide range of fold states with only 3.4% (43/1276) of the peptides being IDPs and 8.2% (105/1276) being TMs. Even though 19.4% (248/1276) of the antimicrobial peptides were not predicted with any fold state, only 4.6% (59/1276) were localized in the free-of-fold-state region revealing that the antimicrobial peptides present quite heterogenous structural properties (Table S5.1 and Supplemental Fig. S5.1). However, it is interesting to note that all three categories present low number of non-predicted cases in the free-of-fold-state region when compared with the unevolved IGORFs (16.8%) supporting that this region may be depleted in functional peptides. All these results correspond to preliminary observations made on three datasets of peptides with functional annotation. A more detailed analysis on multiple categories of functional peptides should be

conducted in order to investigate properly the potential fate of different peptides based on their structural properties.

Perspectives

In section 3, we proposed the IGORFs as potential structural bricks that can serve as starting points for de novo gene emergence or can be combined, thus participating in the overall fold evolution of proteins. Our results highlighted that foldable de novo genes were born through the combination of multiple IGORFs with different foldability potentials supporting that less foldable IGORFs could be integrated into an existing protein fold and profit from the already established structural stability. As a result, we can speculate that IGORFs localized in the region free-of-fold-state, even though highly unstable and prone to aggregate in solution, they could still play an important role in the evolution of folds as elementary structural bricks that could be integrated into an already existing protein, thus stabilizing with the hydrophobic environment provided by the protein. They resonate with short protein fragments, reported so far, that result from different protein structure decompositions with the aim of partitioning protein structures into universal basic units (Berezovsky et al. 2000, 2001; Lamarine et al. 2001; Papandreou et al. 2004; Alva et al. 2015; Postic et al. 2017; Nepomnyachiy et al. 2017). Studying the fold state predictions together with the repartition on the PCA plot of these fundamental protein fragments, would be a very interesting perspective of our analysis which could further describe the fold potential of proteins' elementary structural bricks outside their protein environment. Computational simulations of these small structures (i.e., molecular dynamics) together with experimental validation of their structural state would complete the analysis by shedding more light to their fold stability in solution. These would permit us to study in detail the variability of fold states existing in known protein building blocks and compare it with the one of the potential peptides harboring in the yeast noncoding genome.

6 General Conclusions

The wide use of transcriptomics has revealed a high level of pervasive transcription of presumed noncoding regions, and a fraction of the resulting RNAs have been shown to be translated by ribosome profiling experiments. In addition, mass spectrometry experiments conducted on different organisms confirm the existence of these translation products in the cell with the identification of hundreds of peptides derived from noncoding regions. On the other hand, many studies report examples of de novo gene emergence from noncoding regions. These genes display clear regulation patterns, encode functional proteins and were shown to be subjected to selective pressure. All these results attribute a central role to the so-called noncoding genome in the emergence of genetic novelty, which upon pervasive translation offers the raw material for selection. However, the mechanism behind the emergence of de novo genes stays unclear until today.

The aim of my thesis was to explore the potential role of the noncoding genome in the emergence of genetic novelty and more precisely to investigate how the noncoding genome participates in the emergence of de novo genes as well as in the evolution and structural diversity of proteins. Adopting a structural point of view, I aimed at estimating the potential of the noncoding genome at producing elementary structural bricks which could either serve as the starting points for the birth of de novo genes or be integrated into pre-existing proteins.

The noncoding genome contains the elementary building blocks of proteins

Using the Hydrophobic Clusters Analysis, I showed that the IGORFs of *S. cerevisiae* contain elementary building blocks of proteins. These elementary blocks correspond to clusters of strong hydrophobic amino acids that have been shown to be associated with regular secondary structures (Bitard-Feildel et al. 2018; Lamiable et al. 2019). The HCA clusters of peptides encoded by IGORFs present statistically similar sizes with the ones of the CDS proteins. This result remains true for all the proteins of *S. cerevisiae* no matter if they emerged recently or earlier in the evolutionary time, thereby reinforcing the concept of hydrophobic clusters as elementary building blocks of proteins which are found

widespread in noncoding sequences and are retained along the evolution. In contrast, CDS are enriched in long linkers reflecting long flexible regions on the proteins. Although CDS sequences are becoming continuously longer with the evolutionary time, groups of older proteins present similar linker sizes among them and longer than the ones of the IGORFs. Notably, yeast proteins encoded by young Saccharomyces TRGs and dubious genes present intermediate linker sizes between IGORFs and older proteins, thus supporting that the size of the linkers is a property fixed early in protein evolution.

Study of their amino acid composition revealed that the CDS HCA clusters and linkers are enriched in polar and charged residues compared to those of IGORFs with particularly, negatively charged residues being over-represented. In addition, CDS sequences present higher GC content compared to the IGORFs. Multiple tests on random sequences with different sizes as well as different nucleotide and amino acid compositions revealed that the size of linkers results from a subtle combination of sequence length, GC content, and finally, of the resulting amino acid composition. Nevertheless, it is still unclear whether the linker size is a neutral consequence of the increase of CDS size and enrichment in hydrophilic residues or is a criterion that has been selected along with sequence length and hydrophilic content.

IGORFs encode for peptides that display a wide diversity of fold potential

Moreover, using the HCA foldability score, I showed that IGORFs encode peptides that display a wide range of fold potential diversity including a substantial number of foldable peptides. My fold state prediction model enforced this observation by predicting one (or more) fold state(s) for 72.2% of the potential peptides encoded by IGORFs, highlighting a large range of different fold states harboring in the yeast noncoding genome. In addition, my results on the foldability potential of proteins with different ages support that the yeast proteome tends to evolve towards more foldable proteins supporting that foldability is an important trait that is constantly optimized during evolution. As a matter of fact, globular proteins are usually characterized by a stable and well folded structure which is known to be a requirement for many aspects of their function (Edwards et al. 2013). However, the example of de novo emerged protein Bsc4 supports that young proteins present less optimized folds which do not resemble globular proteins but rather adopt a more rudimentary structure. Notably, Bsc4 presents intermediate linker sizes

between IGORFs and CDS sequences (7.7, 6.3 and 11.5 residues on average, respectively), its hydrophobic residues content is higher than the one of CDS and typical of foldable IGORFs and is significantly depleted in negatively charged residues compared with older proteins. In line with the previous paragraph, it is still unknown whether the linker size is simply the consequence of the enrichment of CDS in hydrophilic residues and the increase in protein size or whether harboring long linkers is accompanied by an increase in foldability and is thus a selected criterion. Overall, all my results highlight an intimate relationship between sequence length, GC content and amino acid composition, whose combination is directly related to the size of linkers and clusters and finally to the foldability of the resulting product. Which one or which combination has driven the evolution of CDS? It is a very complicated question to which we do not have a clear answer yet.

How is the pervasive translation tolerated by the cell?

My results highlight that an important fraction of IGORF-encoded peptides displays high HCA scores and aggregation propensities, thus posing important concerns about the impact of these IGORFs on the cell if they ever become expressed. Notably, my prediction model assigned the majority of these peptides with a transmembrane helix fold state supporting that they may "safely" locate in membranes as proposed in Vakirlis et al. (2020a). Proteins encoded by smORFs have been shown to be localized at the membranes of the cells presenting a broad range of functions such as participate into larger membrane protein complexes or act as protein regulators (Orr et al. 2020). Concerning the rest of high HCA score IGORFs, we can hypothesize that if produced, most of the time, their concentration will not be sufficient so that they become deleterious (Langenberg et al. 2020). Our hypothesis is supported by our observation that lowly abundant proteins are more permissive to higher aggregation propensities than the highly abundant ones (see section 3). As a matter of fact, the results from ribosome profiling experiments show that the translation of IGORFs is most of the times an occasional phenomenon which is not expected to lead to the production of peptides in high concentration. Only a small subset of IGORFs (31 in this study) presents more important signatures of translation and thus a more systematic expression. Interestingly, my fold state prediction model revealed that these IGORFs are enriched in peptides which can acquire a stable structure autonomously (S3) or upon interaction with another protein (DIBS) and depleted in

disordered or transmembrane peptides susceptible to aggregate in the cytosolic environment. Whether the structural properties observed for these highly translated IGORFs favor the emergence of a functionality and the optimization of their translational activity is a question that needs to be further investigated.

On the other hand, even though IGORFs with intermediate HCA scores may exhibit a certain propensity for aggregation we hypothesize that their balanced equilibrium of hydrophobic and hydrophilic residues could permit them to form small compact structures which are either stable in solution or get stabilized through oligomerization or interactions with other proteins. My fold state prediction model supports this hypothesis, as IGORFs with intermediate HCA scores are enriched in peptides expected to be stable in solution or peptides presenting less typical fold properties that get stabilized through disulfide bridges, when compared to the other two HCA score categories. It is interesting to note that an important fraction of these IGORFs could not be assigned with any fold state supporting that they do not display clear similar structural properties with any of the defined fold states. Whether these IGORFs, despite non predicted with a fold state, fold into a specific 3D structure, a partially ordered structure or more dynamic folds that stabilize through oligomerization or embedded into a protein environment (if they are fused with a larger protein) deserves further investigation.

Toward de novo genes

In this study, I propose the IGORFs as elementary modules of novel protein birth and evolution. However, how noncoding sequences become coding is a very difficult question to answer. In order to address this question, I studied the sequence and structural properties of different stages that precede the emergence of de novo genes. Using multiple ribosome profiling data, I identified IGORFs presenting different levels of translation signals. It must be noted that, although we assume that IGORFs with an important signal of translation are more prone to be the starting points which will further give birth to novel genes, this is not the case for all the translated IGORFs as their majority will probably be short-lived and will never become a gene. In addition, using reconstruction methods, I was able to reconstruct the ancestral noncoding sequence preceding the emergence of 70 yeast de novo genes and identify ancIGORFs which indeed gave birth to de novo genes. It should be mentioned that the ancestral reconstruction

methods are based on hypothetical models of evolution and are highly dependent on the quality of the initial sequences' alignment. As a result, whether the reconstructed noncoding sequences correspond to the actual ancestral sequences which gave birth to the de novo genes, we cannot know it with certainty. The observed continuum of structural properties between IGORFs and CDS, recalls the proto-gene model proposed by Carvunis et al. (2012) which suggests that de novo genes emerge from transitory nongenic sequences exhibiting intermediate properties between non-gene and genes. Even though the relation of our highly translated and ancIGORFs with the proto-genes is highly plausible, the continuity between our different intermediate states (translated IGORFs and ancIGORFs) needs to be further demonstrated. Whether IGORFs with high translation signals derive from occasionally translated IGORFs that have optimized their translational activity and whether IGORFs that gave birth to de novo genes are related with highly translated IGORFs are interesting questions which demand further investigations.

The "LEGO brick" model

Studying the relation between the structural properties of ancIGORFs and the ones of their de novo genes, we proposed a model which gives a central role to IGORFs in de novo gene emergence and to a lesser extent in protein evolution. Based on this model, IGORFs can constitute starting points for de novo gene emergence or can be combined, thus increasing protein sizes, and leading to more complex protein architectures. This model unifies two evolutionary processes that are usually addressed separately: the origin of novel genes and the elongation and thus evolution of pre-existing proteins, considering IGORFs as elementary structural bricks widespread in noncoding regions. Our model is further supported by our results showing that proteins become longer and acquire more HCA clusters with the evolutionary time while, at the same time, the sizes of their clusters remain invariant. We can hypothesize that this is the outcome of the combination of consecutive IGORFs that contain clusters of similar sizes.

Moreover, systematic study of the ancIGORFs and their corresponding de novo genes revealed that ancIGORFs with different foldability potentials could be combined and give birth to overall foldable de novo genes. This suggests that newly integrated IGORFs benefit from the structural properties of the preexisting IGORF network and permits the

integration of disordered or aggregation prone IGORFs. Whether the extreme structural properties of these peptides produced as monomers in the cytosol would be tolerated by the cells or whether they are more likely to "exist" being integrated into protein structures is a very interesting question.

Predicting the "behavior" of peptides in the cell with machine learning

During my thesis, I developed a supervised machine learning model in order to better characterize the fold state of the potential peptides encoded by IGORFs. The advantage of this model is that it is trained on amino acid sequences corresponding to real experimentally characterized structures of short peptides and consequently can give us an indication about the potential behavior of the peptides in the cell. Overall, the model presented good performance at predicting the fold state of independent datasets of coding sequences. Further study of the properties of the five different fold state categories with PCA reveals a continuum among them which reflects their overall hydropathy. Disordered peptides enriched in polar residues are clearly separated from highly hydrophobic transmembrane peptides while the other three fold states lie in-between. However, I identified a region along this continuum which does not correspond to any of the defined fold states. Whether this region free of fold state corresponds to a fold category that we did not consider in our five fold states of reference (technical bias) or it reflects a real gap in the structural space (biological bias) is a very interesting question. Therefore, a systematic and exhaustive research of different types of peptides from multiple sequence and structural databases is needed in order to identify potential fold states that are missing from this initial analysis.

However, almost 17% of the total IGORFs were not predicted with a specific fold state and located in this region free-of-fold-state. Based on their high hydrophobic content (though not as high as the TMs) we can stipulate that if produced, they would correspond to peptides unstable in polar environments and which may stabilize only under oligomerization or embedded in mixed polar-apolar environments such as the membrane surface. One should notice that proteins correspond also to those mixed environments presenting a highly hydrophobic core and a polar surface. In line with our "LEGO brick" model, is tempting to speculate that these hydrophobic peptides, unable to fold as monomers in aqueous environments (and potentially deleterious), could participate in

the proteins' construction benefiting from the overall protein environment. Multiple methods of protein structure decomposition into universal basic units have been proposed (Berezovsky et al. 2000, 2001; Lamarine et al. 2001; Papandreou et al. 2004; Alva et al. 2015; Postic et al. 2017; Nepomnyachiy et al. 2017). An interesting perspective would be to investigate whether some elementary structural bricks of proteins can locate in this free-of-fold-state region thus, supporting that these peptides are underrepresented in the nature as monomers but can exist inside protein structures.

Concerning the fold state predictions on the potential peptides encoded by IGORFs, my model assigned one (or more) fold state(s) for 72.2% of the IGORFs supporting that the IGORFs can acquire a wide range of fold states, in line with the results of the HCA fold potential score. However, the PCA plot revealed that multiple IGORF-encoded peptides predicted as TM, S3 or Small fold states were located outside their corresponding ellipses towards the region free-of-fold-state. Notably, my results on known folded peptides revealed that although real TM peptides were systematically located outside their corresponding ellipse, the model was capable at recognizing them and correctly predicting them as TM fold state. These observations pose questions about the accuracy of the predictions made by my model and therefore, its overall accuracy. Do these IGORFs predicted as foldable in solution (S3 and Small) really correspond to the fold state they are assigned to, even though located out of their corresponding ellipse (like the example of the TMs), or is the model wrong when assigning them to a fold state? This is a question to which I do not have a clear answer.

However, I must highlight that my prediction model was trained on a dataset of fold categories extracted from the coding world. These peptides are evolutionary optimized in order to satisfy the structure-stability-function compromise and consequently these evolutionary constraints are expected to have shaped the amino acid composition of their sequences. On the contrary, the IGORFs correspond to unevolved sequences which may be able to fold but have not been optimized for a function and therefore their amino acid sequences present a more important variability. As a result, I am posing the question to which extend we can apply a model trained to recognize fold states corresponding to coding sequences to unevolved noncoding sequences. This relates with the observations made on the fold prediction of AlphaFold[2] on the proteome of *S. cerevisiae*. As a matter of

fact, AlphaFold[2] has been trained to predict folds that are well represented by structures in the PDB database while it struggles to predict with accuracy the 3D structure of young proteins that are probably characterized by more rudimentary folds and are underrepresented in the structural databases.

On the other hand, my model did not make a fold state prediction for 27.8% of the IGORFs. 60% of the non-predicted cases populate the free-of-fold state regions while the remaining 40% correspond to peptides located inside the ellipses of different fold states. It is unclear why my model is not capable at assigning them to a specific fold category, however I can hypothesize that they could potentially acquire a 3D fold in solution according to their localization on the PCA. It must be also mentioned that the PCA is a dimensionality-reduction method which permits to visualize highly complex datasets in simple 2D or 3D scatterplots. Inevitably, this data transformation into new uncorrelated variables (PCs) leads to partial loss of information. As a matter of fact, the three first PCs of the PCA are capable at explaining up to 70.4% of the overall data variance. Studying the repartition of the data beyond the three first PCs, could potentially clarify the reason why these cases were not predicted as a fold state, even though they overlap with fold state ellipses at the first three PCs. In addition, the PCA was used as an initial representation of the data with the aim to capture general tendencies. A more careful and detailed comparison of the descriptors among the different predictions made by my model is of crucial importance to understand their subtle differences.

Finally, both these observations, the mislocation of some fold state predictions from their corresponding ellipses and the overlap of non-predicted cases with fold state ellipses, make us pose question about the initial dataset on which the model was trained. Whether the training set consisted of too specific examples of the predefined fold states thus, making the model to face difficulty at generalizing in cases with more variability? Although the redundancy of homologous sequences (more than 30% identity) was eliminated from the training set, this still stays a possible scenario. Enlarging the training set and monitoring the impact of the variability in it, is to be further investigated.

Methodologies developed during this thesis

During my thesis I developed a package of bioinformatic tools named ORFmine which is available via Github: https://github.com/i2bc/ORFmine. ORFmine consists of two bionformatic tools developed by me and my colleague Nicolas Chevrollier in python3. The first one, named ORFtrack, aims at "tracking" all the ORFs of a genome and annotate them based on their overlapping with known annotated genomic features. This tool was used in order to extract all the IGORFs from the noncoding genome of *S. cerevisiae*. The second one, named ORFold aims at predicting the foldability potential together with the disorder and aggregation propensity of any amino acid sequence by making use of three independent bioinformatic tools: pyHCA, IuPRED2 and Tango, respectively. It was used in order to estimate the fold potential of the IGORFs and detect their HCA clusters and linkers. Both these bioinformatic tools are independent but their combination can provide a global picture of the fold potential and the structural properties of all the potential peptides of a genome. ORFtrack can be applied to any sequenced and annotated genome (from bacteria to human) while ORFfold can be used to any amino acid sequence longer than 20 residues. In addition, with the help of a master student who I supervised, Camille Rabier, I developed another tool, named ORFribo, which aims at mapping correctly ribosome profiling data on the IGORFs of a given genome. This tool was used in order to identify interesting IGORF candidates presenting weak or high signals of translation (occasionally and highly translated, respectively). ORFribo, is not yet integrated in the ORFmine package as it still needs some adjustments. Finally, I developed a supervised SVM-based machine learning model which aims at predicting the fold state of a given short amino acid sequence and thus complement the fold potential estimation of ORFold. Even though in a preliminary state, this model presents good performance at predicting the fold state of known structured peptides. However, its predictions on peptides encoded by unevolved IGORFs are debatable and more controls need to be done.

Perspectives

During my thesis I developed a pipeline which permitted me to study the overall fold potential and the structural properties of all the potential peptides encoded by the noncoding genome of *S. cerevisiae*. The yeast was used as model organism grace to its normal sized and well annotated genome. Once my protocol is set, it would be interesting to apply it on different organisms with different level of genomic complexity and thus

explore the repertoire of fold potential harboring in their noncoding genomes. Some preliminary results on different organisms (*E. coli, H. volcanii, D. melanogaster, M. musculus*) were presented in the Methodology section but more systematic analyses should be conducted towards the exploration of the noncoding world.

In addition, the recent breakthrough of AlphaFold2 and its high performance at predicting the 3D structure of proteins has revisited the subject of protein folding based on the advancement of the deep learning approaches. Although our results support that AlphaFold2 predicts with lower accuracy young proteins with less canonical folds, it would be interesting to study its prediction performance on the ensemble of the potential peptides encoded by the IGORFs and compare them with the ones of the fold potential predicted by HCA. Whether IGORFs that present intermediate HCA scores are predicted with higher accuracy, thus supporting their overall foldable character is a very interesting question that should be further explored. In addition, other methods rather than AlphaFold2 exist, aiming at predicting the 3D fold of proteins or peptides with *ab initio* approaches rather than deep learning (Maupetit et al. 2009; Yang et al. 2020). Predicting the 3D structure of peptides encoded by IGORFs with different fold prediction approaches, would permit us to generate multiple repertoires of structural models of the IGORFs. Comparison of these structural models would permit us to identify IGORFs presenting similar structure predictions with different methods (permitting us to be more confident about their 3D structure prediction) from IGORFs presenting more diverse structure predictions.

Moreover, the translation of IGORFs is an occasional event that cannot be captured easily by ribosome profiling experiments which correspond to a genome-wide snapshot of the cell's translation. Consequently, the detection of these noncanonical translation events demands the integration of multiple ribosome profiling data in order to increase the probability of detecting this weak signal of translation and differentiate it from experimental noise. During my thesis, I combined the information of five independent ribosome profiling experiments and detected IGORFs with weak and stronger signal of translation. Nowadays, multiple ribosome profiling data are stored on public databases and their vast information still stays unexploited. Expanding the number of experiments (and maybe of experimental conditions) and combining their information, would permit

us to characterize with more confidence the translational activity of the noncoding genome and identify interesting IGORF candidates presenting more systematic translation. ORFribo is an automated method permitting to correctly map Ribo Seq reads on the noncoding genome of an organism. It could be used in order to create a large and interactive database which would integrate the information of multiple ribosome profiling data.

7 Supplemental Methods

Estimation of the fold potential, the aggregation, disorder and TM propensities

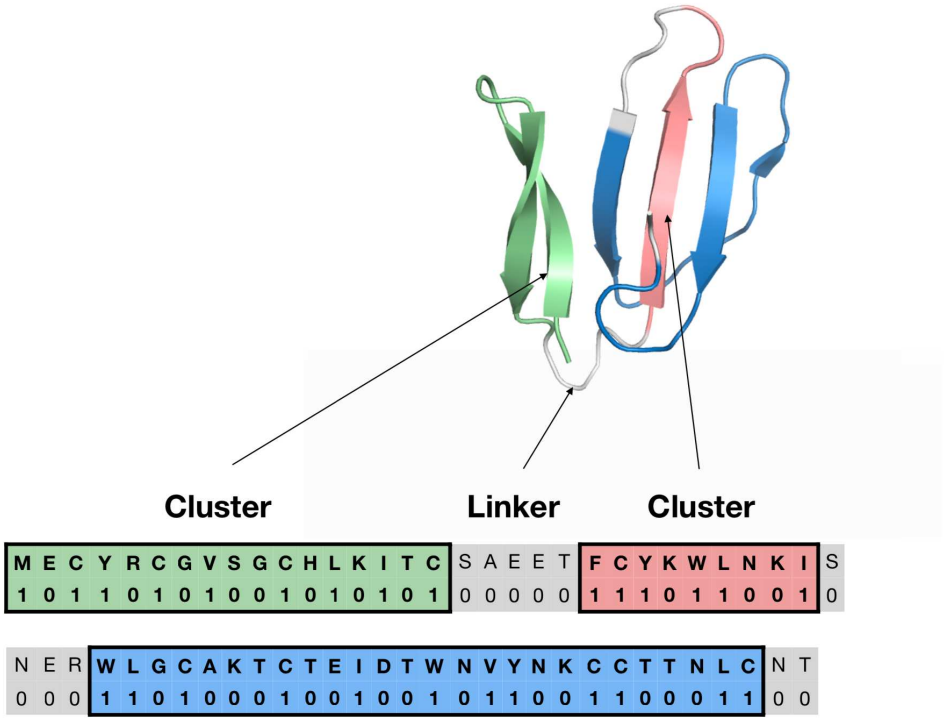
The foldability potential was estimated using a score derived from the HCA (Hydrophobic Cluster Analysis) approach using the HCAtk (Bitard-Feildel and Callebaut 2018a; Bitard-Feildel et al. 2018). HCA divides a protein sequence into (i) clusters gathering strong hydrophobic residues (V, I, L, F, M, Y, W) or cysteines, and (ii) linkers composed of at least 4 non-hydrophobic residues (or a proline). The fold potential of a sequence is determined by its density in hydrophobic clusters but also by the density of hydrophobic amino acids within these clusters. It is reflected with the HCA score which ranges from -10 to +10 where low and high HCA scores indicate sequences which are likely to be disordered or expected to form aggregates in solution. respectively. The aggregation propensity of a sequence was assessed with TANGO (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006a). Following the criteria presented in Linding et al. (2004), a residue was considered as participating in an aggregation prone region if it was located in a segment of at least five consecutive residues which were predicted as populating a b-aggregated conformation for more than 5%. Then, the aggregation propensity of each sequence is defined as the fraction of residues predicted in aggregation prone segments. The disorder propensity was probed with IUPred (Dosztanyi et al. 2005; Mészáros et al. 2009; Dosztányi 2018; Mészáros et al. 2018) using the short prediction option. To be consistent with the criteria used for assessing the aggregation propensity, we considered a residue as participating in a disordered region if it is located in a segment of at least five consecutive residues, each presenting a disorder probability higher than 0.5. Then, the disorder propensity of each sequence is defined as the fraction of residues predicted in disordered prone segments.

Ribosome Profiling analyses

Ribosome profiling experiments: Cells were grown overnight in 0.5 liter of liquid glucose-YPD till an OD_{600} of 0.6, 50 microg/microl of cycloheximide were added to the culture and incubated during 5 min and kept at + 4°C. The pellet of yeast cells was recovered by centrifugation during 5 min at 5000 rpm in Beckman F10 rotor at + 4°C. Total RNA and polysomes were extracted as previously described (Baudin-Baillieu et al. 2014). Briefly, cells were lysated by vortex during 15 min in 500 microl of polysome

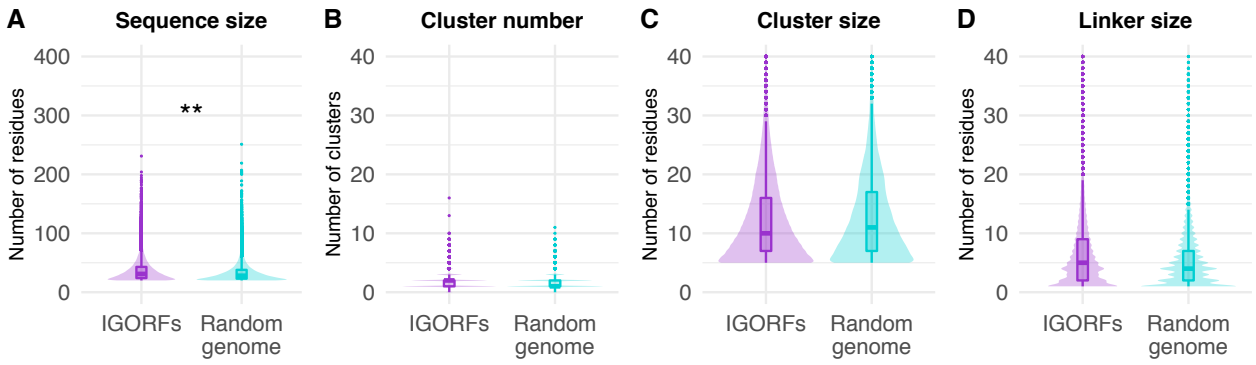
buffer (10 mM Tris-acetate pH7.5 ; 0.1M NaCl and 30 mM Mg-acetate) in presence of glass beads in Eppendorf tube, followed by 5 min of centrifugation at 16 krcf at + 4°C. Ribosome-protected mRNA fragments (RPFs) were generated by the treatment following the ratio of 1 OD_{260nm} of extract with 15 U of RNase I during 1 h at 25°C. Monosomes were collected by 2h15 min centrifugation on a 24% sucrose cushion at +4°C on TLA 110 rotor at 110 krpm. The monosomes were resuspended with 500 microl of polysome buffer. RNA was purified by phenol–chloroform extraction and 28-34 nucleotides RPFs were recovered by electrophoresis in a 17% acrylamide (19/1) 7M urea in 1x TAE gel. These RPFs were depleted of ribosomal RNA by treatment with the Ribo-Zero Gold rRNA removal kit for yeast from Illumina company. RPF libraries were generated with NEBNext Small RNA Sample Prep Kit, according to the manufacturer's protocol, and were checked with the bioanalyser small RNA kit. Sequencing was performed by a HighSeq 2000 (Illumina) 75-nucleotide single-read protocol.

8 Supplemental material



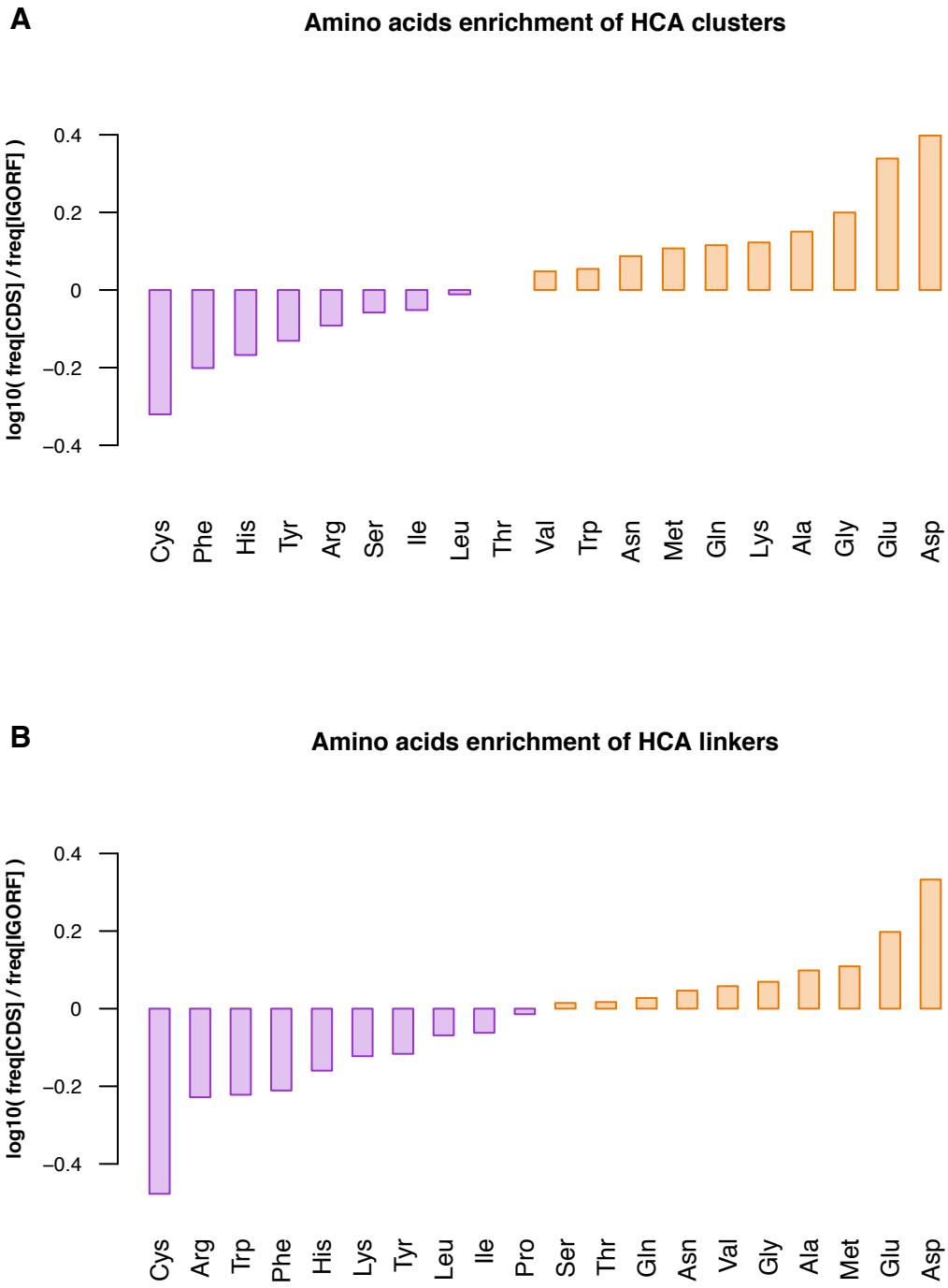
Supplemental Figure S3.1 | 3D mapping of HCA hydrophobic clusters and linkers

HCA hydrophobic clusters (colored) and linkers (in grey) delineated for the sequence of Bucandin (pdb code: 1f94). The HCA-based sequence, which consists in translating the protein sequence into a binary pattern, is given under the protein sequence. “1” corresponds to strong hydrophobic amino acids (V, I, L, F, M, Y, W) and “0” to the other amino acids (see Methods for more details). HCA clusters and linkers are mapped on the 3D structure of Bucandin with respect to the color code used for the sequence.



Supplemental Figure S3.2 | Random IGORFs behave similarly to real IGORFs for most properties

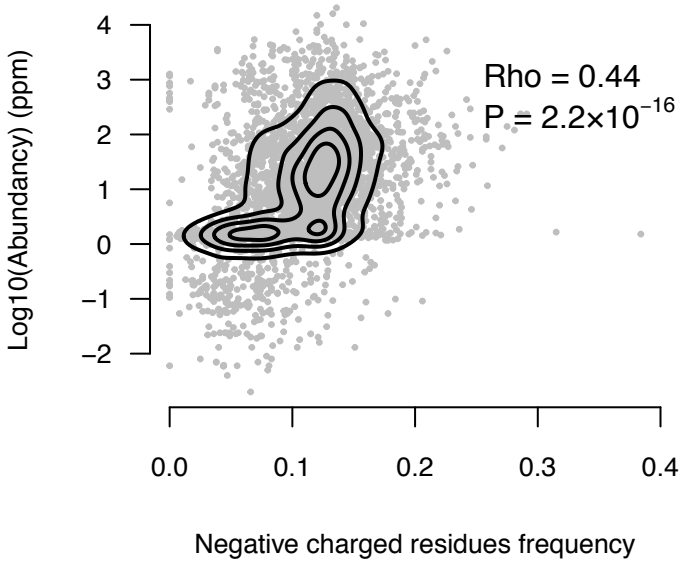
Boxplot distributions of sequence and HCA-based structural properties of real IGORFs and random IGORFs (A) sequence size (B) number of HCA clusters per sequence (C) size of HCA clusters (D) size of linkers. Asterisks denote level of significance: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$



Supplemental Figure S3.3 | CDS are enriched in hydrophilic residues

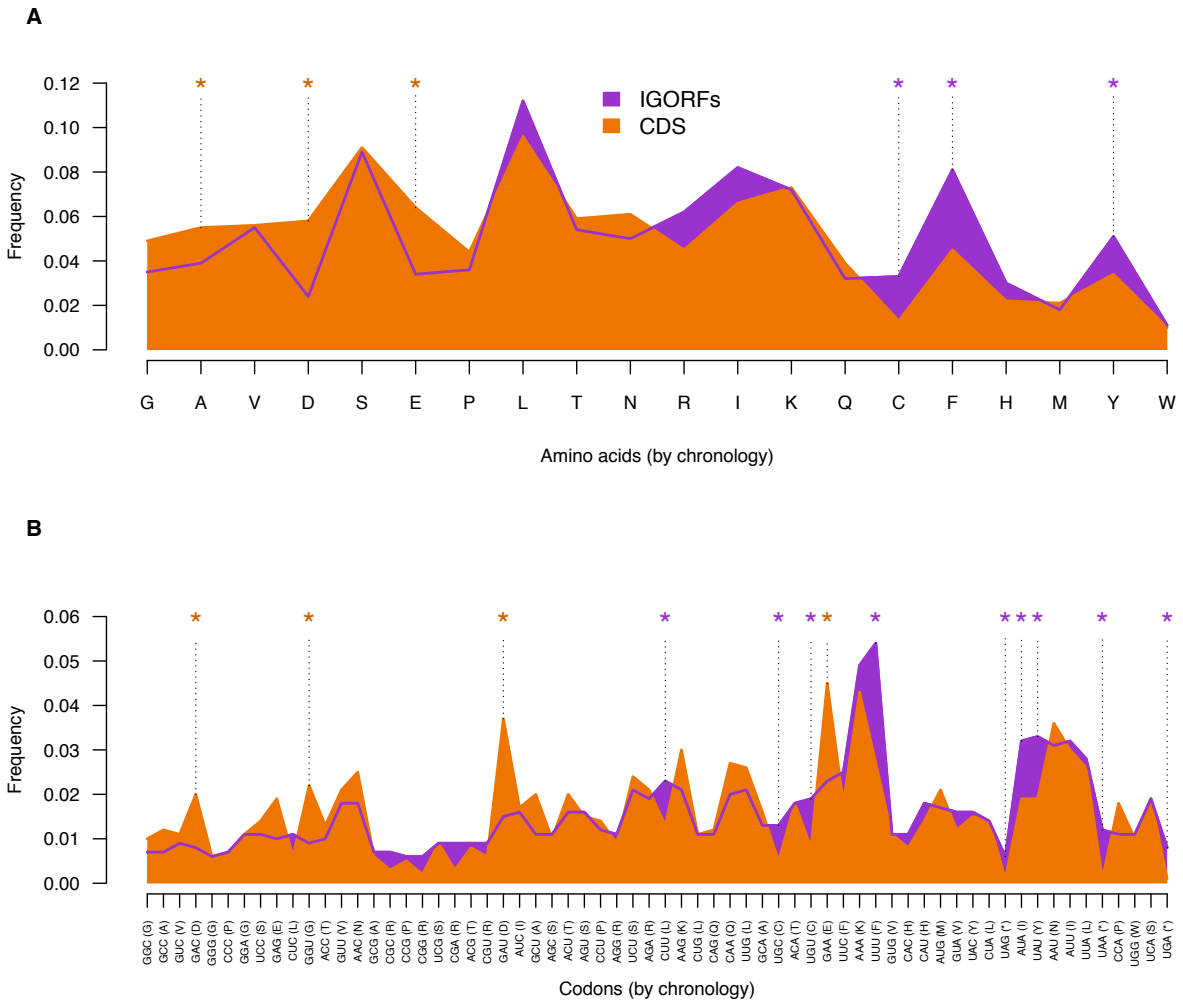
(A) Log ratios of amino acid frequencies in HCA clusters of CDS versus HCA clusters of IGORFs. Negative values (purple) correspond to amino acids with higher frequency in IGORF HCA clusters while positive values (orange) correspond to amino acids that are more frequent in CDS HCA

linkers. (B) Log ratios of amino acid frequencies in HCA linkers of CDS versus HCA linkers of IGORFs.



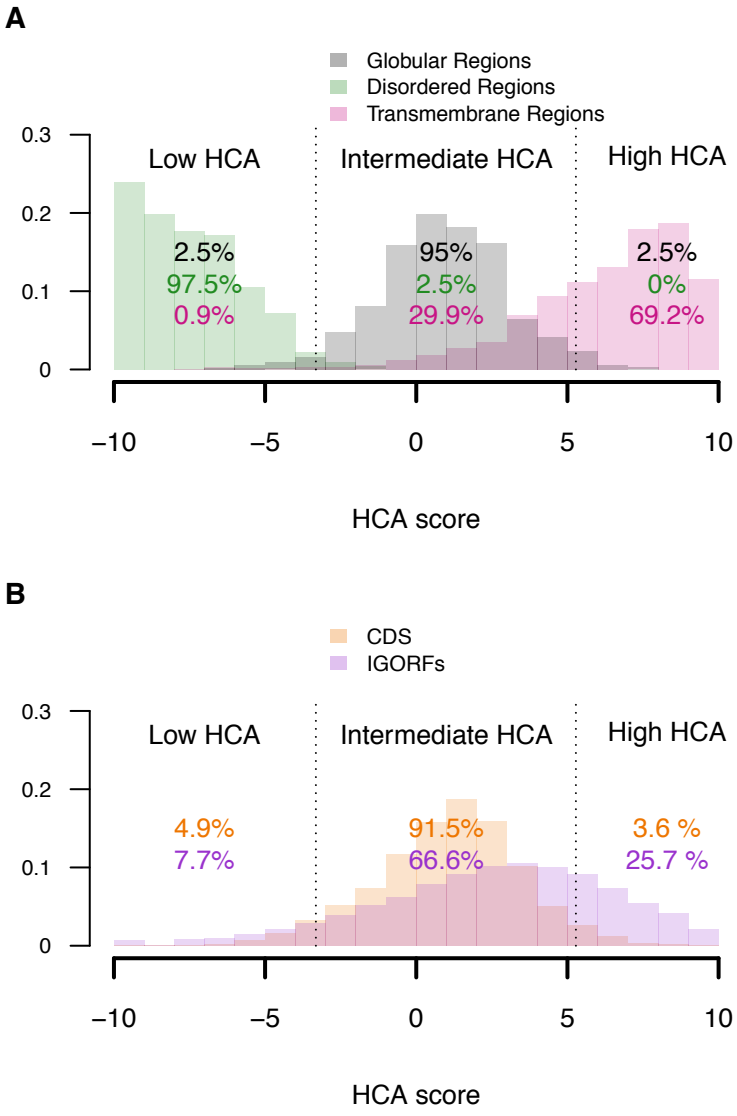
Supplemental Figure S3.4 | Abundant proteins are enriched in negatively charged amino acids

Protein abundances (in parts per million) of all cytoplasmic proteins are plotted against their corresponding negatively charged residues (Aspartate and Glutamate) frequencies. The Spearman rank correlation coefficient is indicated on the plot (p-value < 2.2×10^{-16}).



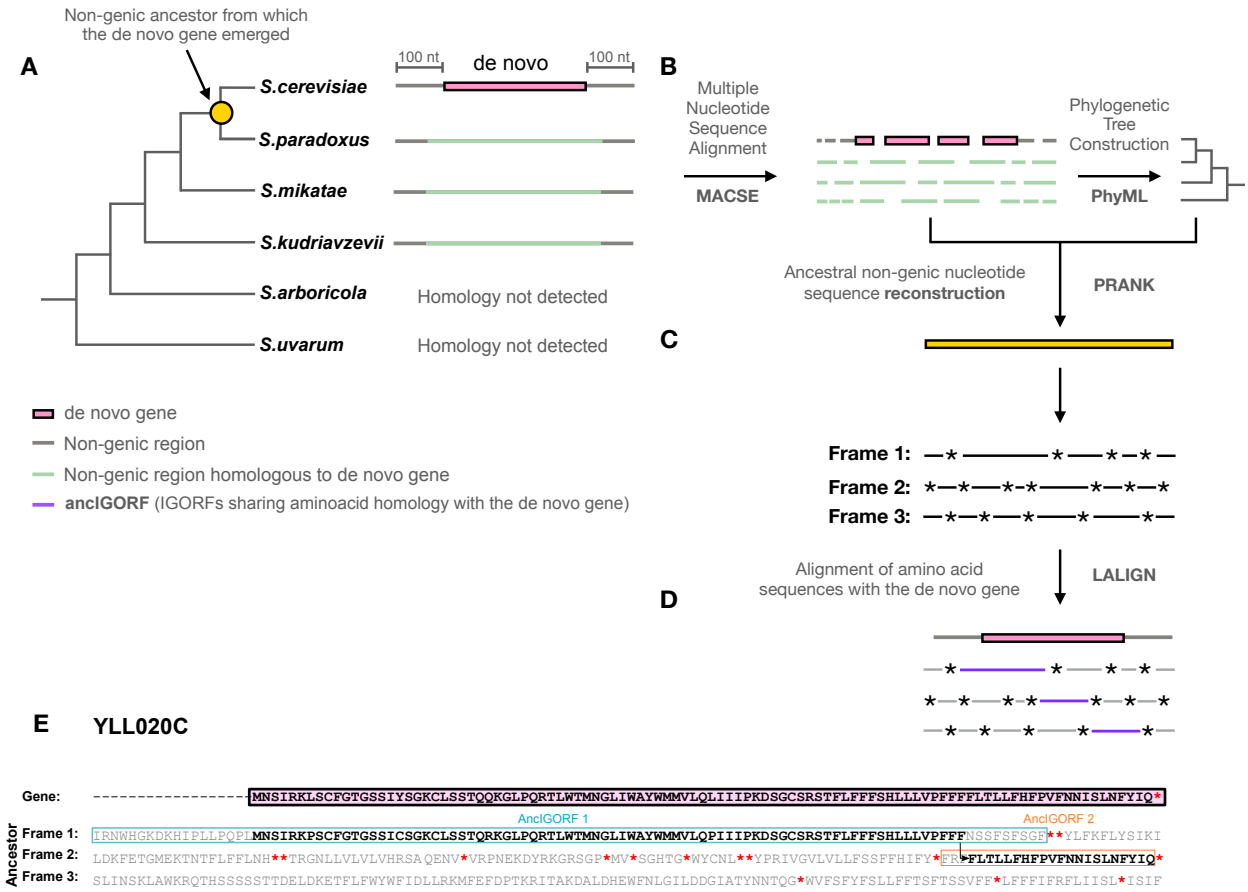
Supplemental Figure S3.5 | CDS are enriched in ancient amino acids

(A) Frequencies of amino acids of CDS (orange) and IGORFs (purple) ordered according to their chronology of appearance during evolution as defined in Trifonov et al. (2001) (B) Frequencies of codons of CDS (orange) and IGORFs (purple) ordered according to their chronology of appearance during evolution as defined in Trifonov et al. (2001). Amino acids or codons enriched in CDS or IGORFs are indicated by orange or purple stars respectively (z-test, p-values < 5×10^{-2}).



Supplemental Figure S3.6 | IGORFs encompass the large spectrum of fold potential of canonical proteins (raw data)

(A) Histograms of the HCA scores of the three reference datasets (i.e. disordered regions, globular domains and transmembrane regions – green, black and pink histograms respectively). Dotted black lines delineate the boundaries of the low, intermediate and high HCA score categories. The boundaries are defined so that 95% of globular domains fall into the intermediate HCA score category whereas the low and high HCA score categories include all sequences with HCA values that are lower or higher than those of 97.5% of globular domains respectively. (B) Histograms of the HCA scores of CDS and IGORFs. The percentages of sequences in each category are given for all datasets.



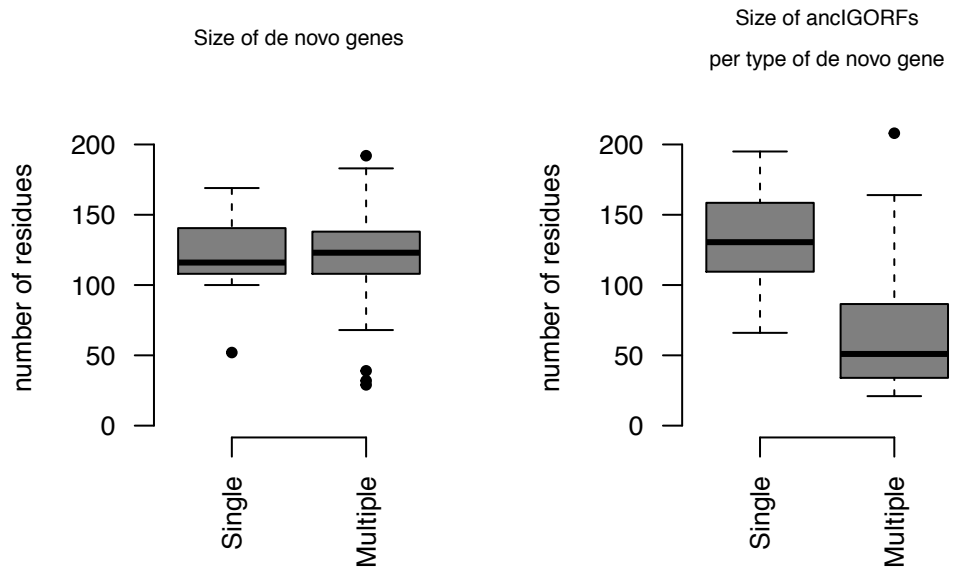
Supplemental Figure S3.7 | Reconstruction of the ancestral IGORFs (ancIGORFs) which gave birth to known de novo genes

(A) Identification of homologous sequences (that can be an orthologous gene or a homologous noncoding sequence) of the de novo gene of interest in all neighboring species with blast (Altschul et al. 1990) (see Methods for more details) (B) Multiple sequence alignment of the detected homologous nucleotide sequences with MACSE (Ranwez et al. 2011, 2018) and construction of their phylogenetic tree with PhyML (Guindon et al. 2010) (C) reconstruction of the corresponding ancestral nongenic nucleotide sequence (in yellow) with PRANK (Löytynoja and Goldman 2010). The latter is subsequently translated into the three frames. STOP codons are indicated with stars. (D) Alignment of all the reconstructed IGORFs (amino acid sequences) with the de novo gene(s) of interest with LALIGN (Huang and Miller 1991) and detection of the IGORFs sharing a homology with it (i.e. ancIGORFs) (E) Alignment of the *S. cerevisiae* de novo gene YLL020C with the translation products of its corresponding ancestral noncoding sequence as predicted for the ancestor of *S. cerevisiae* and *S. paradoxus*. STOP codons are indicated with red stars. The two IGORFs which gave birth to the YLL020C gene (ancIGORFs) are indicated by blue and orange boxes respectively. The two ancIGORFs are distributed across two frames showing that the current version of YLL020C results from a frameshift event. The sections of the ancIGORFs that

participate in the resulting de novo gene are indicated in bold. The HCA scores of the blue and orange IGORFs are 0.48 (foldable) and 7.71 (aggregation-prone) respectively.

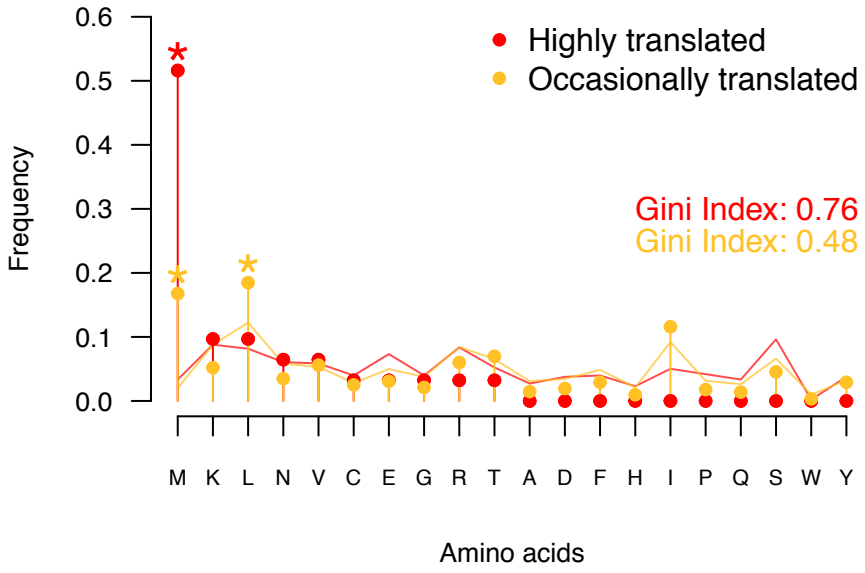
Supplemental Figure S3.8 | Appearance of a Methionine and fusion of two ancIGORFs in the *S. cerevisiae* lineage

The sequences of the YOR333C de novo gene and its corresponding noncoding regions in the five neighboring species of *S. cerevisiae* are indicated in blue. The ancestral sequences are indicated in yellow. STOP codons are represented with red stars. The appearance of the Methionine in the *S. cerevisiae* lineage is highlighted with a grey box while the STOP codon mutation that led to the fusion of the two ancIGORFs in the *S. cerevisiae* lineage is indicated with a green box.



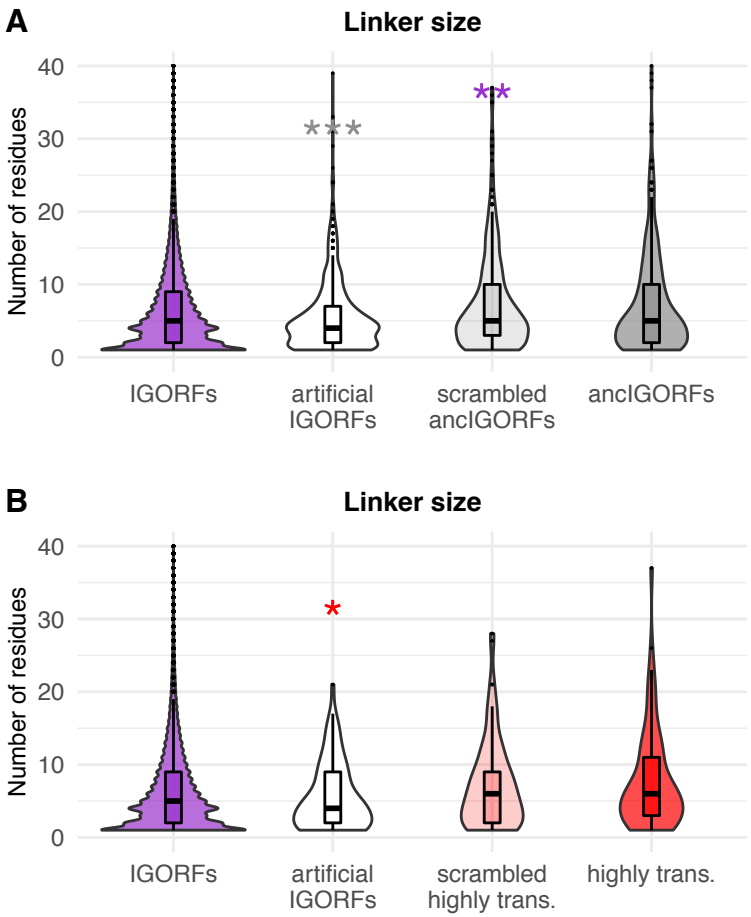
Supplemental Figure S3.9 | De novo gene categories display similar sizes while their corresponding ancIGORFs exhibit different sizes

(A) Boxplot comparing the sequence size of multiple and single ancIGORF de novo genes. (B) Boxplot comparing the sequence size of ancIGORFs preceding the emergence of single and multiple ancIGORF de novo genes.



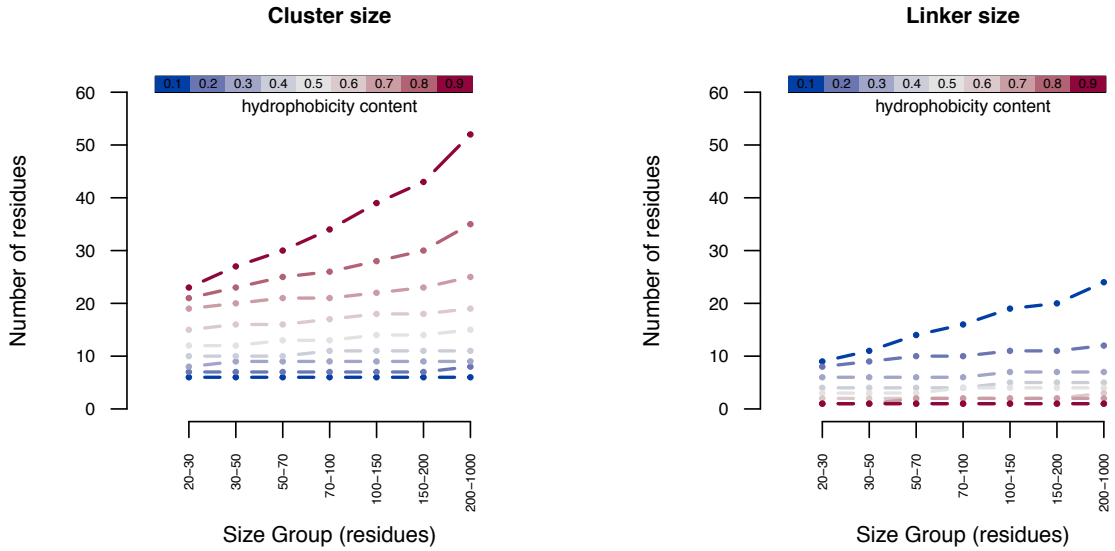
Supplemental Figure S3.10 | Translated IGORFs are mostly initiated with Methionine

Frequencies of the 20 amino acids at the first translated position for highly translated IGORFs (red) and occasionally translated ones (yellow). Gini indexes which reflect the statistical dispersion of the 20 amino acids at the first translated position are given for highly and occasionally translated IGORFs in red and yellow respectively. Gini index values range from 0 to 1 and high values reflect the fact that the first translated positions are enriched in specific amino acids, particularly, in MET and to a lesser extent in LEU for occasionally translated IGORFs. Amino acids which are significantly observed at the first translated position compared to the other translated positions are indicated with a star (z-test p.value < 5×10^{-2}).



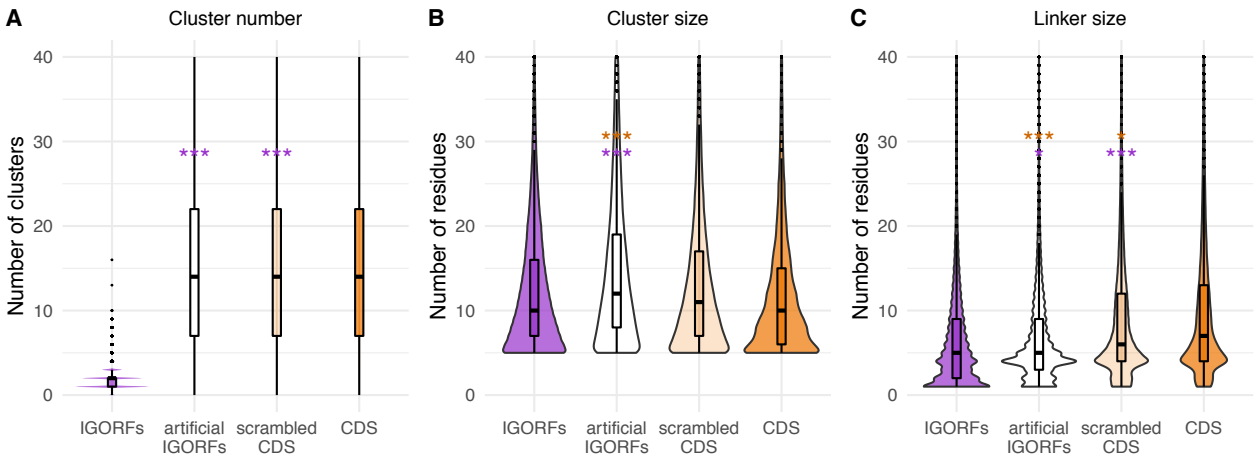
Supplemental Figure S3.11 | The nucleotide composition of ancestral and highly translated IGORFs seems to play an important role in the linker's size

(A) Linkers' size for real IGORFs (purple), artificial IGORFs (i.e. ORFs with size similar to ancIGORFs but nucleotide composition of IGORFs) (white), ancIGORFs with scrambled nucleotides (light grey) and real ancIGORFs (grey). (B) Linkers' size for real IGORFs (purple), artificial IGORFs (i.e. ORFs with size similar to highly translated IGORFs but nucleotide composition of IGORFs) (white), highly translated IGORFs with scrambled nucleotides (ligh tred) and real highly translated IGORFs (red). The p-values were computed with the Mann-Whitney *U* test (one-sided). Asterisks denote level of significance: *$p < 5 \times 10^{-2}$, **$p < 1 \times 10^{-2}$, ***$p < 1 \times 10^{-3}$. The color of the asterisks indicates the ORF category used for the comparison.



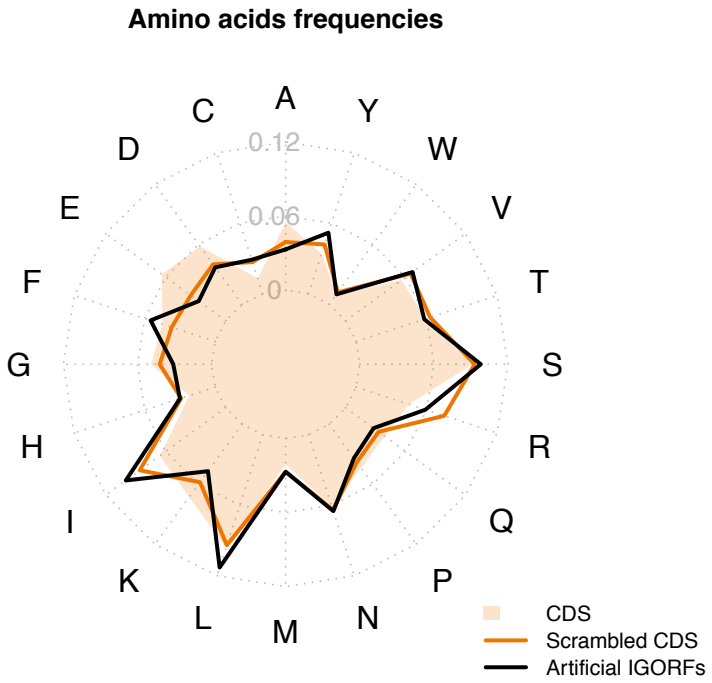
Supplemental Figure S3.12 | Impact of the hydrophobicity content and sequence length on the size of clusters and linkers

In order to properly decipher the contributions of the amino acid composition and sequence length, we generated artificial sequences with different sizes and different hydrophobic residue contents (1000 sequences per bin of sequence size and hydrophobicity content). (A) The median values of the resulting cluster sizes are subsequently plotted in number of residues. (B) For the same artificial sequences, the median values of the resulting linker sizes are plotted in number of residues. In both plots sequences are colored according to their hydrophobicity content that ranges from 0.1 (i.e. 10% of strong hydrophobic residues according to HCA definition: V, I, L, M, Y, F, W and C) to 0.9. For a given sequence length, hydrophobic and hydrophilic contents have a significant impact on the size of clusters and linkers respectively with an even more important effect on long sequences.



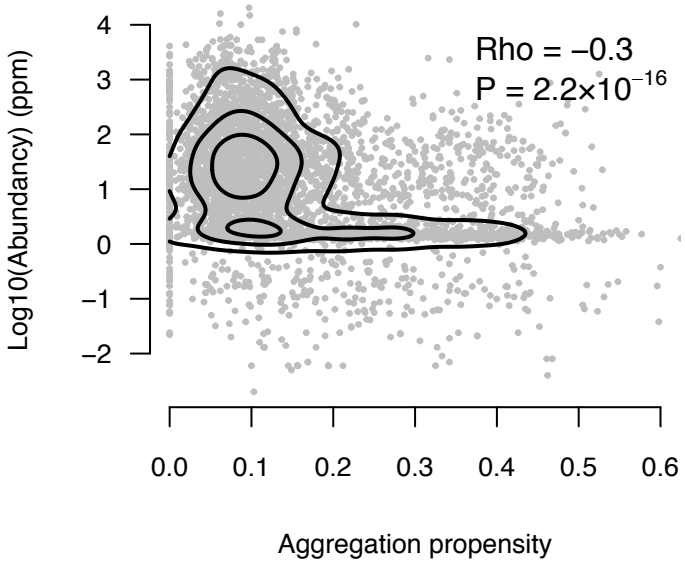
Supplemental Figure S3.13 | Effect of the sequence length, and GC content on the size of clusters and linkers

Number of HCA clusters (A), size of HCA clusters (B) and size of linkers (C) for real CDS sequences (orange), scrambled CDS sequences (light orange) and artificial IGORFs (i.e. with size similar to CDS but nucleotide compositions of IGORFs (white). The clusters of scrambled CDS are similar to those of CDS while their linkers are slightly shorter (Mann-Whitney *U* test, $P = 4 \times 10^{-2}$) showing that randomly and according to the GC content and size of CDS, long though slightly shorter linkers can be generated. In contrast, the linkers of artificial IGORFs are of comparable size to those of IGORFs though slightly larger, while the artificial clusters are longer (Mann-Whitney *U* test, $P = 4 \times 10^{-2}$ and $P = 6 \times 10^{-4}$ respectively). This reflects that at the IGORF GC content, the sequence length alone has a small impact on cluster size while the effect is marginal on linker size, and overall cannot explain the increase in linker size observed for CDS. Indeed, the artificial linkers are clearly shorter than those of both real and scrambled CDS (Mann-Whitney *U* test, $P = 7.1 \times 10^{-8}$ and 2×10^{-4} respectively) highlighting the impact of the amino acid composition but also of the GC content of the CDS on their linker size. The p-values were computed with the Mann-Whitney *U* test (one-sided). Asterisks denote level of significance: $*p < 5 \times 10^{-2}$, $**p < 1 \times 10^{-2}$, $***p < 1 \times 10^{-3}$. The color of the asterisks indicates the ORF category used for the comparison.



Supplemental Figure S3.14 | Impact of the GC content on the resulting amino acid compositions

Radar plot reflecting the 20 amino acid frequencies for real CDS (light orange shadow), scrambled CDS (orange line) and artificial IGORFs (i.e. sequences with size similar to CDS but nucleotide compositions of IGORFs (black line)). CDS and artificial IGORFs exhibit slightly different GC contents (GC content of 36.1% and 39.6% for IGORFs and CDS respectively) that lead to slightly different amino acid compositions.

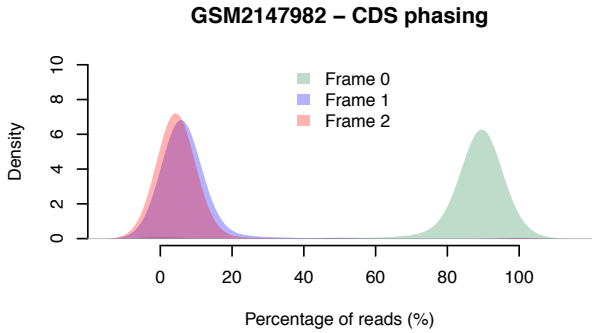


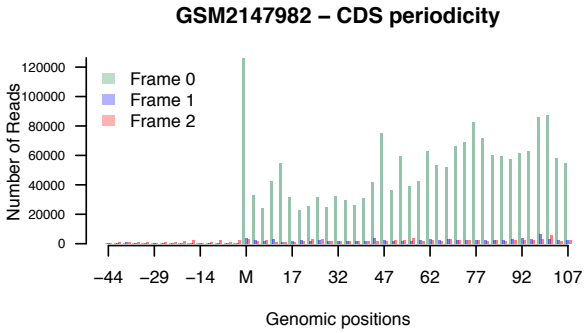
Supplemental Figure S3.15 | Lowly abundant proteins display a large spectrum of aggregation propensities

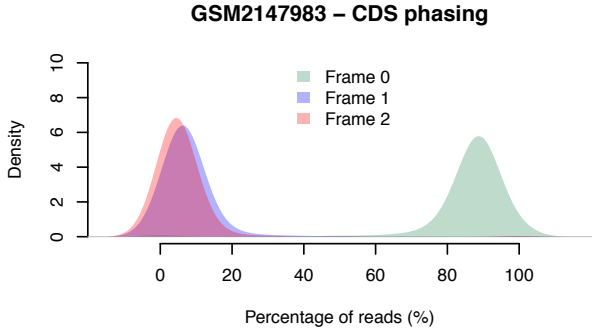
Protein abundances (in parts per million) of all cytoplasmic proteins are plotted against their corresponding aggregation propensity predicted with TANGO (Linding et al. 2004; Fernandez-Escamilla et al. 2004; Rousseau et al. 2006). The Spearman rank correlation coefficient is -0.30 with p-value < 2.2×10^{-16}.

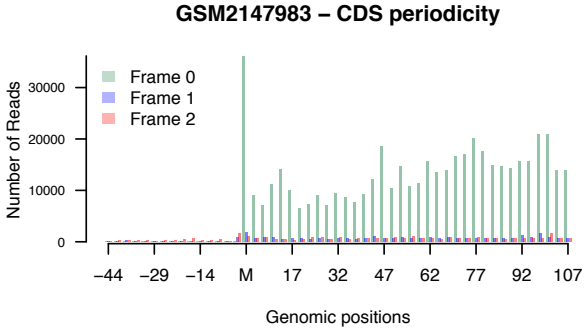
Supplemental Figure S3.16 | The fusion of IGORFs can lead to longer clusters or linkers

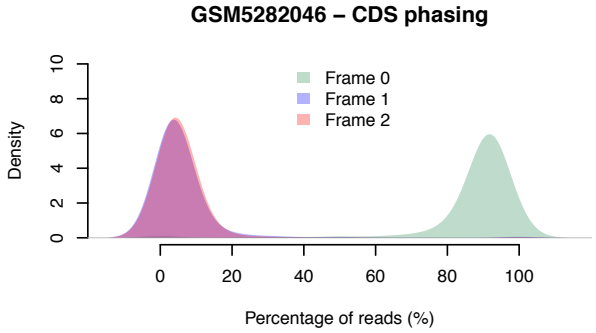
The sequence of the YMR153C-A de novo gene (A) and YPR126C (B) are indicated by the blue boxes while their corresponding ancestral sequences are indicated by the yellow boxes. STOP codons are represented by red stars. HCA clusters are highlighted by red boxes while HCA linkers correspond to the regions connecting two HCA clusters or extremities that are not associated with an HCA cluster.

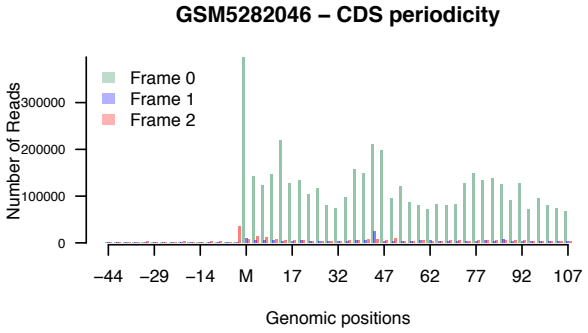

GSM2147982 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
0 2 4 6 8 10
0 20 40 60 80 100
Percentage of reads (%)

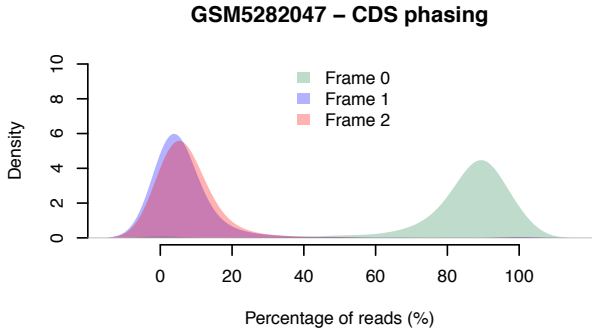

GSM2147982 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
0 20000 40000 60000 80000 100000 120000
−44 −29 −14 M 17 32 47 62 77 92 107
Genomic positions

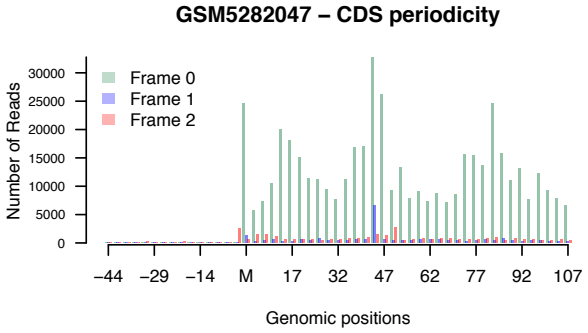

GSM2147983 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
0 2 4 6 8 10
0 20 40 60 80 100
Percentage of reads (%)

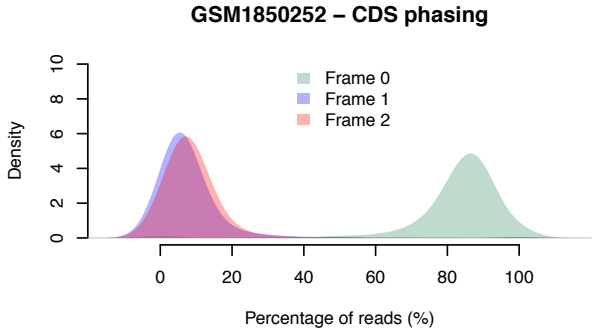

GSM2147983 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
0 10000 20000 30000
−44 −29 −14 M 17 32 47 62 77 92 107
Genomic positions

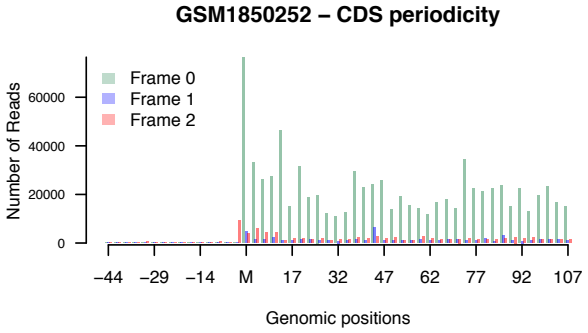

GSM5282046 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
0 2 4 6 8 10
0 20 40 60 80 100
Percentage of reads (%)


GSM5282046 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
0 100000 200000 300000
−44 −29 −14 M 17 32 47 62 77 92 107
Genomic positions


GSM5282047 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
0 2 4 6 8 10
0 20 40 60 80 100
Percentage of reads (%)


GSM5282047 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
0 5000 10000 15000 20000 25000 30000
−44 −29 −14 M 17 32 47 62 77 92 107
Genomic positions


GSM1850252 – CDS phasing
Density
Frame 0
Frame 1
Frame 2
0 2 4 6 8 10
0 20 40 60 80 100
Percentage of reads (%)


GSM1850252 – CDS periodicity
Number of Reads
Frame 0
Frame 1
Frame 2
0 20000 40000 60000
−44 −29 −14 M 17 32 47 62 77 92 107
Genomic positions

Supplemental Figure S3.17 | Quality control for the 28-mer RPFs used for the detection of occasionally and selectively translated IGORFs for all five experiments

The left panel shows that 90% (in average) of the 28-mer RPFs are in frame with the start codon of the CDS (Frame 0). The right panel presents the number of RPFs at each nucleotide position (determined by the site P of each 28-mer) showing accumulation of signal over the CDS (reads detected only after the start codon), and a nice periodicity (of frame 0) over the 100 first nucleotides. Both these results inform us about the good quality of the RPF data in all five experiments.

Supplemental Table S3.1. One-sided Mann-Whitney *U* test p-values for all the ORF categories – Sequence length (in amino acids)

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	3×10^{-4}	3×10^{-2}	2.2×10^{-23}	5.2×10^{-38}	1.4×10^{-153}
Occasionally translated		-	2×10^{-1}	1.3×10^{-15}	1.8×10^{-36}	1.0×10^{-150}
Highly translated			-	1.3×10^{-3}	1.5×10^{-13}	2.2×10^{-19}
Ancestral IGORFs				-	1.1×10^{-16}	1.7×10^{-63}
De novo genes					-	5.2×10^{-20}
CDS						-

Supplemental Table S3.2. One-sided Mann-Whitney *U* test p-values for all the ORF categories - Number of clusters

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	2×10^{-2}	6×10^{-2}	3.2×10^{-15}	1×10^{-35}	7×10^{-148}
Occasionally translated		-	2×10^{-1}	8×10^{-11}	3.3×10^{-33}	1.9×10^{-142}
Highly translated			-	1×10^{-2}	1.1×10^{-10}	2.2×10^{-18}
Ancestral IGORFs				-	1.7×10^{-13}	1.8×10^{-60}
De novo genes					-	3.3×10^{-20}
CDS						-

Supplemental Table S3.3. Two-sided Mann-Whitney *U* test p-values for all the ORF categories – Cluster size

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	5×10^{-1}	7×10^{-1}	6×10^{-1}	8×10^{-2}	1×10^{-1}
Occasionally translated		-	6×10^{-1}	6×10^{-1}	6×10^{-2}	1×10^{-1}
Highly translated			-	7×10^{-1}	4×10^{-1}	2×10^{-1}
Ancestral IGORFs				-	5×10^{-2}	1×10^{-1}
De novo genes					-	2×10^{-3}
CDS						-

Supplemental Table S3.4. One-sided Mann-Whitney *U* test p-values for all the ORF categories – Linker size

	IGORFs	Occasionally translated	Highly translated	Ancestral IGORFs	De novo genes	CDS
IGORFs	-	1×10^{-1}	2×10^{-2}	1×10^{-2}	9×10^{-5}	6.3×10^{-11}
Occasionally translated		-	9×10^{-2}	1×10^{-1}	2×10^{-3}	1.5×10^{-8}
Highly translated			-	7×10^{-1}	3×10^{-1}	8×10^{-3}
Ancestral IGORFs				-	3×10^{-2}	7.9×10^{-7}
De novo genes					-	1.1×10^{-3}
CDS						-

Supplemental Table S3.5. Strong hydrophobic residues (V,I,L,F,M,Y,W) frequency per ORF category for the three HCA score categories.

	Low HCA	Intermediate HCA	High HCA	Total
IGORFs	0.239	0.391	0.508	0.410
Occasionally translated	0.241	0.384	0.494	0.401
Highly translated	0.251	0.355	0.406	0.353
Ancestral IGORFs	0.241	0.376	0.508	0.392
De novo genes	0.215	0.398	0.479	0.410
CDS	0.219	0.332	0.475	0.328

Supplemental Table S3.6. The 70 de novo genes of *Saccharomyces cerevisiae* used for the ancestral reconstruction. For the last two columns the hydrophobic residues considered are: V,I,L,F,M,Y,W and the hydrophilic ones are: K,R,D,E,Q,N.

Gene name	Ancestral type	Protein size	HCA score	HCA bin	Clusters count	Disorder propensity	Aggregation propensity	Hydrophobic percentage	Hydrophilic percentage
YAL026C-A	multiple	145	5.54	high	5	0	0.407	0.475	0.31
YAL031W-A	multiple	102	3.87	intermediate	4	0	0.284	0.431	0.225
YAL047W-A	single	109	4.06	intermediate	6	0.055	0.312	0.451	0.203
YBL100W-C	multiple	39	5.97	high	3	0.205	0.256	0.386	0.361
YBR056C-B	single	52	-5.44	low	1	0.5	0.096	0.211	0.308
YBR206W	multiple	107	1.68	intermediate	4	0.056	0.121	0.299	0.243
YCL058C	single	152	6.28	high	6	0	0.546	0.52	0.107
YCR085W	multiple	117	4.33	intermediate	6	0	0.385	0.512	0.196
YDL016C	single	100	0.78	intermediate	3	0	0.05	0.37	0.25
YDL158C	multiple	102	7.32	high	2	0	0.598	0.509	0.256
YDR024W	multiple	161	0.01	intermediate	7	0.062	0.174	0.336	0.273
YDR154C	multiple	116	0.34	intermediate	3	0.086	0.19	0.379	0.198
YDR327W	multiple	108	1.98	intermediate	5	0.046	0.093	0.381	0.27
YDR396W	multiple	166	4.39	intermediate	8	0	0.373	0.385	0.222
YDR426C	multiple	125	6.5	high	4	0	0.392	0.504	0.232
YER014C-A	multiple	153	4.46	intermediate	10	0.039	0.301	0.399	0.248
YER046W-A	multiple	109	2.96	intermediate	7	0.073	0.165	0.404	0.212
YER076W-A	single	115	3.82	intermediate	3	0.087	0.252	0.409	0.26
YER087C-A	multiple	183	2.54	intermediate	9	0.055	0.213	0.382	0.131
YER133W-A	multiple	113	2.3	intermediate	4	0.071	0.124	0.39	0.239
YFR026C	single	169	1.56	intermediate	6	0.101	0.213	0.314	0.32
YGL152C	multiple	225	5.4	high	9	0	0.409	0.422	0.129
YGL165C	multiple	192	3.51	intermediate	7	0.026	0.349	0.421	0.218
YGL214W	single	161	0.34	intermediate	7	0.05	0.081	0.324	0.267
YGR011W	multiple	108	3.06	intermediate	5	0	0.352	0.407	0.24
YGR050C	multiple	118	1.79	intermediate	5	0.153	0.042	0.372	0.287
YGR064W	multiple	122	5.04	intermediate	5	0	0.115	0.353	0.229
YGR137W	multiple	124	2.85	intermediate	5	0.04	0.298	0.434	0.242
YGR151C	single	111	0.26	intermediate	7	0.045	0.135	0.369	0.387
YHL006W-A	single	117	2.59	intermediate	8	0	0.051	0.352	0.155
YHR022C-A	multiple	29	4.86	intermediate	3	0	0	0.447	0.274
YHR071C-A	single	106	7.51	high	3	0	0.179	0.452	0.264
YHR180W	single	163	1.77	intermediate	9	0.037	0.227	0.404	0.227
YIL028W	multiple	132	4.84	intermediate	6	0	0.333	0.448	0.219
YIL030W-A	multiple	112	5.07	intermediate	4	0	0.384	0.482	0.242
YIL066W-A	multiple	147	2.14	intermediate	6	0.088	0.259	0.341	0.205
YIL071W-A	multiple	158	4.16	intermediate	8	0.082	0.399	0.444	0.196
YIL086C	multiple	102	-0.97	intermediate	5	0.157	0.118	0.332	0.314

YJL077W-B	multiple	32	4.09	intermediate	2	0	0.156	0.436	0.374
YJL119C	single	107	0.29	intermediate	4	0.047	0.224	0.328	0.326
YJL142C	multiple	130	4.83	intermediate	7	0.054	0.523	0.483	0.215
YJL211C	multiple	147	0.02	intermediate	7	0.095	0.136	0.328	0.273
YJR018W	multiple	120	4.38	intermediate	4	0	0.25	0.39	0.184
YJR020W	single	110	3.38	intermediate	5	0	0.245	0.391	0.227
YJR087W	multiple	116	3.87	intermediate	5	0	0.267	0.432	0.207
YKL036C	multiple	130	1.35	intermediate	3	0.131	0.362	0.37	0.261
YKL053W	multiple	124	2.3	intermediate	6	0.056	0.452	0.54	0.217
YKL076C	multiple	127	3.34	intermediate	6	0	0.299	0.447	0.259
YKL123W	multiple	126	4.13	intermediate	5	0	0.31	0.428	0.317
YKL136W	multiple	132	-0.14	intermediate	6	0	0.311	0.355	0.182
YKL153W	multiple	169	3.54	intermediate	6	0.036	0.249	0.414	0.308
YLL020C	multiple	101	3.85	intermediate	6	0.059	0.436	0.487	0.18
YLR041W	multiple	106	4.71	intermediate	3	0.094	0	0.34	0.378
YLR171W	single	129	4.02	intermediate	6	0.047	0.504	0.473	0.165
YLR255C	multiple	117	-1.42	intermediate	3	0.197	0.154	0.308	0.342
YLR412C-A	multiple	68	-4.51	low	2	0.676	0	0.221	0.515
YLR434C	multiple	127	5.39	high	4	0	0.181	0.393	0.329
YMR052C-A	single	121	7.22	high	4	0	0.479	0.562	0.207
YMR103C	multiple	120	-2.13	intermediate	3	0.108	0.158	0.342	0.251
YMR119W-A	single	124	8.48	high	2	0	0.5	0.557	0.216
YMR153C-A	multiple	111	4.12	intermediate	3	0.045	0.252	0.432	0.252
YMR173W-A	multiple	394	2.82	intermediate	21	0.018	0.312	0.445	0.111
YNL150W	multiple	135	1.89	intermediate	6	0.104	0.222	0.348	0.23
YNL226W	multiple	136	2.43	intermediate	5	0.037	0.228	0.434	0.236
YNL269W	multiple	131	1.98	intermediate	7	0.122	0.168	0.375	0.305
YOR316C-A	multiple	69	-1.2	intermediate	3	0.217	0	0.318	0.274
YOR333C	multiple	138	5.46	high	6	0	0.159	0.413	0.326
YPL056C	multiple	101	0.53	intermediate	6	0	0.218	0.367	0.209
YPR126C	multiple	102	7.54	high	4	0	0.461	0.568	0.186
YPR150W	multiple	173	5.63	high	5	0	0.416	0.462	0.169

Supplemental Table S3.7. Frequencies of the three STOP codons for different ORF categories.

	UAA	UAG	UGA
IGORFs	0.45	0.24	0.31
Occasionally translated	0.48	0.23	0.29
Highly translated	0.48	0.32	0.20
CDS	0.47	0.23	0.30

Table S5.1. Fold state predictions made by our model on different datasets of functional peptides.

Prediction	Antimicrobial peptides (%)	Toxic peptides (%)	Cell communication peptides (%)
IDPs	3.4	1.8	17.5
DIBS	22.9	10.0	28.3
S3	11.0	8.0	9.2
Small	22.5	52.1	17.5
TMs	8.2	6.2	1.6
Multiple	12.5	7.4	6.7
Non-Predicted	19.4	14.5	19.2

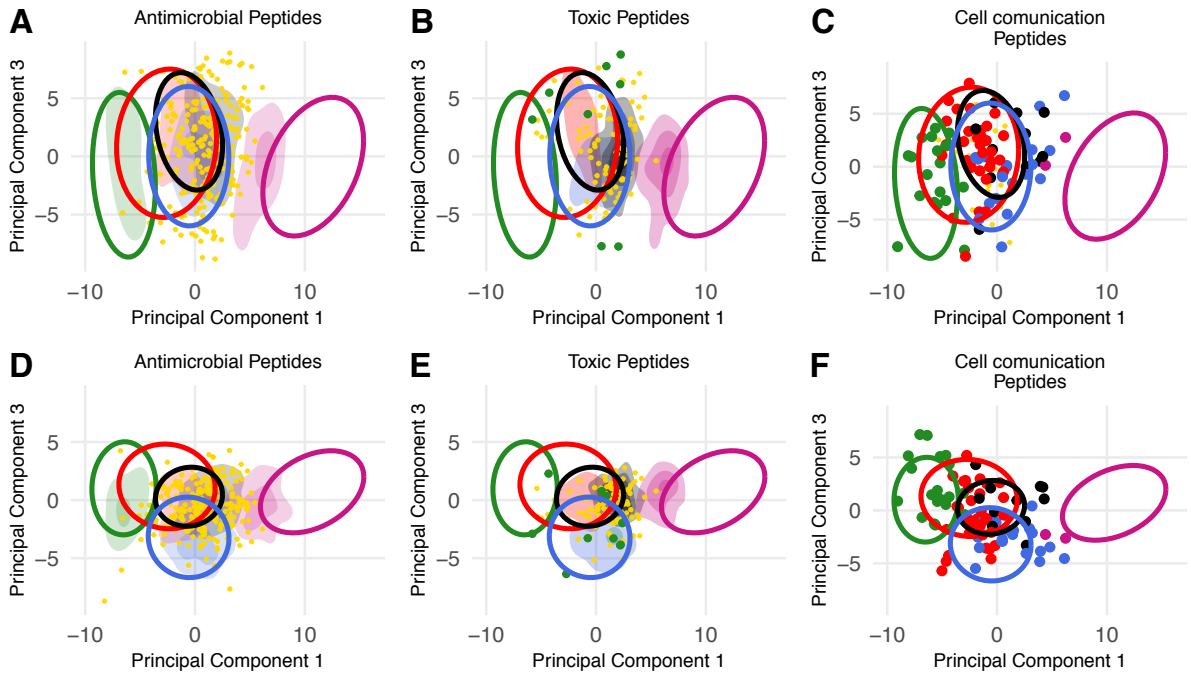


Figure S5.1. Principal Components Analysis of the training set presented with colored ellipses (ellipse level at 90% of the data) and projection of the different categories of functional peptides (in colored densities). **(A)** PCA of the first and the second PCs for antimicrobial peptides; **(B)** PCA of the first and the second PCs for toxic peptides; **(C)** PCA of the first and the second PCs for cell-cell communication peptides; **(D)** PCA of the first and the third PCs for antimicrobial peptides; **(E)** PCA of the first and the third PCs for toxic peptides; **(F)** PCA of the first and the third PCs for cell-cell communication peptides. The non-predicted cases are projected in yellow points. All the data are colored based on the fold state prediction made by our model; Colors: Green for IDPs, Red for DIBS, Blue for Small, Black for S3 and Purple for TMs.

Résumée en français

Le génome non codant

Les efforts pour détecter et annoter les séquences codant pour les protéines (appelées CDS) dans les génomes à l'aide d'approches bioinformatiques se sont traditionnellement appuyés sur des règles arbitraires telles que la conservation des acides aminés des séquences, la recherche d'un codon de départ AUG et une longueur minimale de 50 ou 100 acides aminés (Basrai et al. 1997, Couso et Patraquim 2017, Chen et al. 2020). Ces règles d'annotation ont été largement adoptées pour des raisons de commodité et afin d'assurer un faible nombre de faux positifs. En conséquence, plusieurs petits ORFs (smORF), sans preuve expérimentale de fonction, ont été systématiquement rejetés et pendant de nombreuses années, tous ces smORF intergéniques ont été considérés comme des séquences non-codantes (Basrai et al. 1997; Ruiz-Orera et Albà 2019). Dans les années 1960, le terme « Junk DNA » est devenu très populaire tandis qu'en 1972, le terme a été officiellement inventé indépendamment par Susumu Ohno (Ohno 1972) et David Comings (Comings 1972). Ce « terme provocateur » a été utilisé pour souligner « l'inutilité » de ces régions d'ADN.

Le génome non-codant n'est pas aussi silencieux qu'on le croyait

Les estimations actuelles indiquent que moins de 2 % du génome des mammifères code pour des séquences codantes (Clark et al. 2011 ; Lybecker et al. 2014). Cependant, les approches de transcriptomique ont révélé que la grande majorité du génome est transcrite au-delà des limites des gènes connus (Kapranov et al. 2002 ; Clark et al. 2011). Ce phénomène est défini comme une transcription omniprésente et est responsable de la génération d'un grand ensemble de différentes molécules d'ARN différentes de celles qui codent pour les protéines canoniques et celles avec des fonctions déjà établies telles que les ARNt, les ARNr, les snRNA et les snoRNA (Jensen et al. 2013). Ces observations, de transcription au-delà des gènes codant pour les protéines, ont attribué un rôle essentiellement régulateur au génome non-codant. Il est devenu clair que l'ADN jusqu'à présent appelé « Junk DNA » n'était pas aussi inutile et inactif qu'on le pensait auparavant et qu'il pouvait participer à de multiples fonctions cellulaires.

De plus, ces dernières années, le translatome (i.e. étude de la traduction de l'ensemble d'un génome) de nombreux organismes eucaryotes a été largement exploré et a prouvé qu'en plus des séquences codant pour les protéines longues classiques (gènes codant pour les protéines annotés), il existe de nombreux petits ORFs présumés non-codants qui peuvent être traduits, conduisant à la production de petits peptides à partir de régions génomiques présumées non-codantes (Ingolia et al. 2009, 2011 ; Ruiz-Orera et al. 2018 ; Ruiz-Orera et Albà 2019). Ces smORFs avec des signatures de traduction sont principalement détectés dans des transcrits précédemment considérés comme non-codants. Après leur identification, il a été prouvé que beaucoup d'entre eux ont des rôles régulateurs importants pour l'expression des gènes, pour produire des micropeptides fonctionnels ou même être simplement le résultat de la traduction omniprésente conduisant probablement à la production de peptides non-fonctionnels (Ruiz-Orera et Albà 2019).

Les produits codés par les smORF sont appelés peptides codés par smORF (SEP) ou micropeptides. Pendant des années, les smORF et leurs micropeptides correspondants ont été négligés en raison de leur petite taille et de la difficulté de leur détection (Makarewich et Olson 2017). Cependant, grâce aux études protéiques et à l'avancement des techniques de spectrométrie de masse, plusieurs micropeptides ont été caractérisés et par conséquent, le domaine de la peptidomique a attiré plus d'attention.

Tous ces résultats prouvent qu'un nombre non négligeable de loci en dehors des régions codant pour les protéines bien définies chez les bactéries (Ndah et al. 2017; Weaver Jeremy et al. 2019), les champignons (Ingolia et al. 2009; Wilson et Masel 2011 ; Carvunis et al. 2012), les animaux (Ingolia et al. 2011 ; Chew et al. 2013 ; Bazzini et al. 2014 ; Ingolia et al. 2014 ; Aspden et al. 2014 ; Ruiz-Orera et al. 2018) et les plantes (Hanada et al. 2007 ; Hsu et al. 2016), sont transcrits et traduits de manière omniprésente, conduisant à la production de petites protéines dans la cellule. Des questions intéressantes concernant le devenir de ces petites protéines et leur impact sur la cellule peuvent être posées. Ces peptides pourraient-ils acquérir un rôle fonctionnel pour la cellule, ou sont-ils simplement le résultat d'une traduction non-canonique ?

En fait, malgré leur petite taille, il a été démontré que les micropeptides jouent un rôle essentiel dans de nombreux processus biologiques, notamment le développement, la réparation de l'ADN, l'homéostasie du calcium, le métabolisme, la signalisation du stress, la fusion des myoblastes et la mort cellulaire (Makarewich et Olson 2017). Certains micropeptides, codés par les smORF dans les régions 5' UTR des gènes, jouent souvent un rôle régulateur dans l'expression des gènes, révélant un rôle fonctionnel important de ces petites protéines. En parallèle, les progrès des technologies de séquençage de l'ADN et les puissantes méthodes de séquençage de nouvelle génération ont conduit au séquençage de plusieurs génomes (Hu et al. 2011). La disponibilité de plusieurs génomes séquencés a fait progresser le domaine de la génomique comparative (Gerstein et al. 2007) et a révélé plusieurs exemples de gènes fonctionnels codant pour des protéines qui ont émergé de régions auparavant non-codantes, appelés gènes de novo (Levine et al. 2006; Cai et 2008 ; Li et coll. 2009 ; Knowles et McLysaght 2009 ; Li et coll. 2010 ; Murphy et McLysaght 2012 ; Gubala et coll. 2017 ; Vakirlis et coll. 2018 ; Zhang et coll. 2019a).

Toutes ces études montrent que le génome dit non-codant est un réservoir important de petits ORFs qui, lors d'une transcription et d'une traduction omniprésentes, peuvent produire un nombre important de petits peptides dans le cytosol cellulaire. Néanmoins, la plupart de ces peptides, s'ils ne sont pas délétères pour la cellule, devraient être de courte durée et instantanément dégradés. Cependant, de nombreux exemples montrent que parfois de nouveaux produits fonctionnels peuvent émerger de cette procédure. Tous ces résultats attribuent un rôle central au génome non-codant dans l'émergence de la nouveauté génétique, qui, lors d'une traduction omniprésente, offre la matière première pour la sélection et l'évolution de gènes de novo.

Les gènes de novo

Pendant de nombreuses années, l'émergence de novo de gènes à partir de séquences d'ADN auparavant non-codantes a constitué un événement rarement observé et n'a pas été considéré comme un processus évolutif potentiel de naissance de gènes (Jacob 1977 ; Siepel 2009 ; Ohno 2013 ; Zhang et al. 2019b). Susumu Ohno, dans son livre « Evolution by gene duplication » (Ohno 2013) soutient que tous les nouveaux gènes proviennent de gènes déjà existants tandis que François Jacob a affirmé que « la probabilité qu'une protéine fonctionnelle apparaisse de novo par association aléatoire d'acides aminés est

pratiquement zéro » (Jacob 1977). Cependant, l'émergence de novo à partir de régions non-codantes s'est maintenant avérée être un mécanisme supplémentaire indéniable et des études rapportant des preuves de la naissance de gènes de novo sont publiées chaque année.

Plusieurs modèles ont été proposés pour tenter d'expliquer le mécanisme sous-jacent de l'émergence de gènes de novo. Le modèle du « continuum » repose sur l'hypothèse que les gènes issus de novo pourraient initialement présenter des caractéristiques simples et devenir progressivement plus complexes au cours de l'évolution (Carvunis et al. 2012). Les auteurs ont introduit la notion de proto-gènes qui correspondent à des stades intermédiaires et réversibles de la naissance de gène de novo. D'autre part, contrairement au modèle des proto-gènes, le modèle de préadaptation soutient que les gènes récemment apparus devraient afficher des caractéristiques géniques exagérées, plutôt que des caractéristiques intermédiaires entre les séquences non géniques et géniques (Wilson et al. 2017; Van Oss et Carvunis 2019). Selon Wilson et al. (2017), les nouveaux gènes ne naissent que de séquences qui se trouvent être pré-adaptées, pour ne pas être nocives pour la cellule. De telles séquences non-nocives sont des séquences solubles, avec une forte propension intrinsèque au désordre, leur permettant d'éviter d'agréger ce qui serait délétère dans l'environnement cellulaire aqueux. Un autre modèle, nommé « grow slow and moult », soutient que les ORF codant pour les protéines pourraient éventuellement étendre leurs extrémités via la traduction au-delà des bordures de la séquence codante, occasionnelle dans un premier temps et constitutive plus tard, conduisant à l'expression de nouveaux domaines aux N- et C-terminaux. Ces nouveaux domaines peuvent être bien intégrés dans la structure protéique préexistante et être encore affinés par sélection offrant de nouvelles fonctions à l'ancienne protéine ou séparés par leur ORF d'hébergement conduisant à la création d'un gène de novo.

Tous ces modèles soutiennent que le peptide de novo initial, une fois établi, évoluera davantage vers une protéine plus canonique et bien repliée. En conséquence, ils donnent tous un rôle central au potentiel de repliement des ORFs non-codants dans l'émergence de la nouveauté génétique. Notamment, le génome non-codant peut être vu comme un réservoir d'innovation moléculaire apportant une plasticité génétique aux organismes et leur permettant d'évoluer dans leur environnement.

Le but de cette thèse

L'objectif général de ma thèse est d'étudier le rôle potentiel du génome non-codant dans l'émergence de la nouveauté génétique. En particulier, j'ai cherché à étudier comment le génome non-codant participe à l'émergence de gènes de novo ainsi qu'à l'évolution et à la diversité structurale des protéines. Pour répondre à cette question, j'ai adopté un point de vue structural car il est bien connu que la fonction des protéines est intimement liée à leur structure. Par conséquent, j'ai caractérisé la diversité du potentiel de repliement (propension au désordre, à l'état replié ou à l'agrégation) des séquences d'acides aminés codées par tous les ORFs intergéniques (IGORF) de *S. cerevisiae* afin de (i) estimer le potentiel du génome non-codant pour produire de nouvelles briques protéiques, qui peuvent soit donner naissance à de nouveaux gènes, soit être intégrées dans des protéines préexistantes, participant ainsi à l'évolution et à la diversité de la structure des protéines, et (ii) explorer si la grande diversité structurelle observée dans les protéomes est déjà présente dans les séquences non-codantes, et ainsi étudier la relation, le cas échéant, entre le potentiel de repliement des séquences d'acides aminés codées par les IGORF et la diversité structurale des protéines.

Principaux résultats

La première partie de ma thèse consistait à développer une méthode bioinformatique pour la détection de tous les IGORFs de *S. cerevisiae* et l'estimation du potentiel de repliement des peptides potentiels codés par ceux-ci. Pour cela, nous avons développé un outil bioinformatique appelé ORFtrack qui vise à « extraire » tous les ORFs d'un génome donné et annoter leur chevauchement (ou pas, dans le cas des IGORFs) avec des éléments génomiques annotées. Ensuite, nous avons développé ORFold qui vise à estimer le potentiel de repliement ainsi que la propension au désordre et à l'agrégation d'une séquence d'acides aminés donnée et nous l'avons appliqué sur les peptides codés par les IGORFs. ORFold utilise trois outils bioinformatiques académiques et libres (pyHCA (Faure et Callebaut 2013a, 2013b ; Bitard-Feildel et Callebaut 2018 ; Lamiable et al. 2019), IuPRED2 (Mészáros et al. 2018 ; Dosztányi 2018 ; Erdős et Dosztányi 2020) et Tango (Linding et al. 2004 ; Fernandez-Escamilla et al. 2004 ; Rousseau et al. 2006), respectivement) et donne une indication du potentiel de repliement (ainsi que des

informations complémentaires sur le désordre et la propension à l'agrégation) pour chaque séquence d'acides aminés codés par les ORFs codants et non-codants d'un génome. Ces deux outils bioinformatiques sont regroupés dans un package appelé ORFmine qui est librement accessible via GitHub (GitHub 2021). Le package ORFmine ainsi qu'un protocole détaillé étape par étape d'extraction des IGORFs et de l'estimation du potentiel de repliement des peptides qu'ils encodent ont été présentés dans le livre Methods Molecular Biology dans un issue specific à « Computational Peptide Science » dans le chapitre intitulé "*Exploring the peptide potential of genomes*".

Les résultats produits par ces méthodes nous ont permis de caractériser le potentiel de repliement ainsi que d'autres propriétés de séquence et de structure des peptides codés par les IGORFs. En particulier, nous avons montré que les IGORFs codent pour une grande diversité de peptides potentiels, y compris des peptides à forte propension au désordre ou à l'agrégation, et de manière surprenante, une majorité prédite comme capables de se replier en 3 dimensions (i.e. foldables). De plus, en utilisant l'Analyse des Cluster Hydrophobes (HCA), nous avons montré que les IGORF de *S. cerevisiae* contiennent les briques élémentaires de construction de protéines. Ces briques élémentaires correspondent à des amas d'acides aminés hydrophobes (dits clusters HCA) qui ont été montrés comme correspondant aux structures secondaires régulières (Bitard-Feildel et al. 2018 ; Lamiable et al. 2019). Les clusters HCA identifiés dans les peptides encodés par les IGORFs présentent des tailles statistiquement similaires à ceux trouvés dans les protéines encodées par les CDSs. En revanche, les CDSs sont enrichies en régions riches en acides aminés hydrophiles (dits linkers, qui reflètent de longues régions flexibles) plus longues que celles identifiées dans les IGORFs. On émet l'hypothèse que ces longs linkers contribuent probablement à optimiser les arrangements locaux des structures secondaires, offrent une flexibilité aux protéines et une spécificité dans les interactions protéiques. L'étude de leur composition en acides aminés a révélé que les clusters et les linkers HCA de CDSs sont enrichis en résidus polaires et chargés par rapport à ceux des IGORFs, les résidus chargés négativement étant particulièrement surreprésentés. Cela peut s'accompagner d'une augmentation de la spécificité des repliements et des interactions protéiques grâce à l'optimisation des processus de repliement et d'assemblage (Lumb et Kim 1995).

Nous avons ensuite caractérisé les premiers stades précédant l'émergence des gènes de novo avec deux approches complémentaires (i) la reconstruction systématique des séquences ancestrales non-codantes de 70 gènes de novo de *S. cerevisiae* afin d'identifier les propriétés de séquence et de structure des IGORFs qui ont effectivement donné naissance à des gènes de novo connus et (ii) l'identification d'IGORFs avec un fort signal de traduction par des expériences de ribosome profiling, afin d'étudier des propriétés de séquence et de structure des IGORFs candidats qui pourraient donner naissance à de futurs nouveaux gènes. Dans cette partie de ma thèse, nous avons développé un pipeline qui permet de cartographier correctement les données de ribosome profiling sur des séquences non-codantes afin de détecter des IGORFs avec un signal de traduction. Ce pipeline appelé ORFribo fera bientôt partie du package ORFmine proposant un protocole complet pour (i) la détection et l'extraction des IGORFs, (ii) la prédiction du potentiel de repliement de leurs séquences d'acides aminés, et (iii) la caractérisation de l'activité de traduction des IGORFs.

Nous avons montré que bien que les IGORFs ancestraux affichent une large palette de potentiels de repliement, les IGORFs foldables sont plus susceptibles de donner naissance à de nouveaux gènes et que la plupart des gènes de novo résultent de la combinaison de plusieurs IGORFs avec différents potentiels de repliement. De plus, nous avons distingué les IGORFs qui sont traduits de façon occasionnelle avec un faible signal de traduction de ceux qui affichent une signature de traduction forte et avons montré que bien que les premiers ressemblent aux ORFs non-codants en général, les seconds, avec les IGORFs ancestraux, présentent des propriétés de séquence et de structure intermédiaires entre les IGORFs et les gènes codant pour des protéines. Au total, ces résultats nous permettent de proposer un modèle (présenté sur la Figure R1) qui donne un rôle central aux IGORFs dans l'émergence de novo de gènes et dans une moindre mesure dans l'évolution des protéines, complétant ainsi la large palette des mécanismes d'évolution des protéines comme les événements de duplication, transfert horizontal de gènes, réarrangement des domaines… Ce modèle unifie deux processus évolutifs qui sont généralement abordés séparément : l'origine de nouveaux gènes et l'élongation et donc l'évolution de protéines préexistantes, à travers les IGORFs en tant que modules moléculaires élémentaires répandus dans les régions non-codantes. Tous ces résultats sont présentés dans un article de recherche intitulé « *Intergenic ORFs as elementary structural modules of de novo gene*

birth and protein evolution » et qui a été accepté pour publication dans la revue Genome Research.

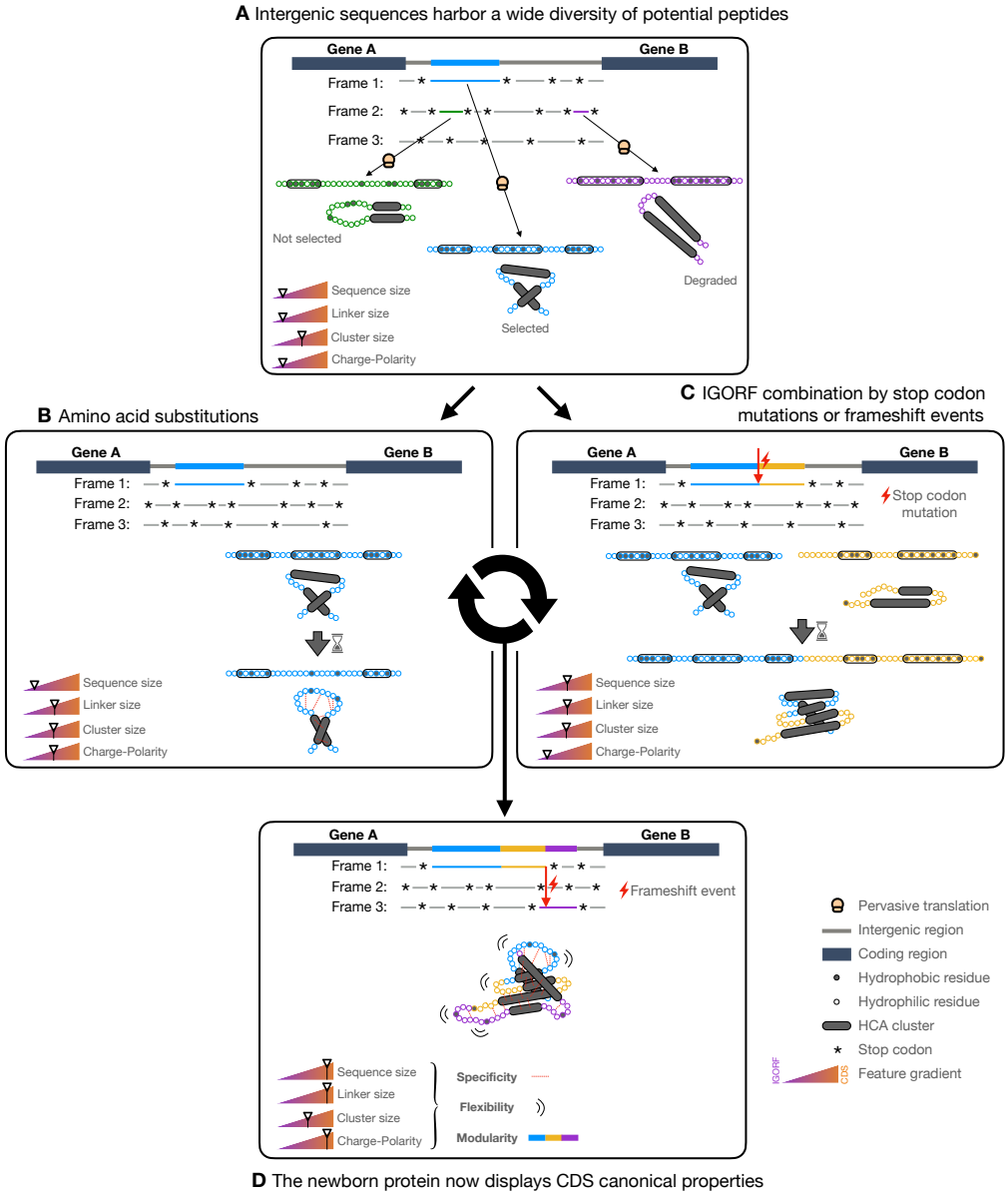


Figure R1. Modèle d'émergence de gènes de novo et d'évolution des protéines avec les IGORFs comme modules structuraux élémentaires. (A) Les IGORFs codent pour une grande diversité de peptides parmi lesquels une grande quantité devrait pouvoir se replier en solution. Lors d'une traduction omniprésente, certains peptides qui peuvent être délétères ou non seront immédiatement dégradés. Parmi les autres, le peptide bleu conférera un avantage à l'organisme et sera sélectionné, fournissant ainsi un point de départ pour la naissance d'un gène de novo. (B)

L'IGORF bleu, une fois sélectionné, est soumis à des substitutions d'acides aminés augmentant ainsi la proportion globale de résidus hydrophiles. Dans le cas présent, cela induit (i) la rupture du second cluster entraînant l'augmentation de la taille du linker central et (ii) l'établissement d'interactions spécifiques entre les résidus hydrophiles (points rouges) qui augmentent la spécificité du processus de repliement. (C) Le codon STOP de l'IGORF bleu peut être muté en un acide aminé, ajoutant ainsi l'IGORF jaune à l'IGORF sélectionné préexistant et allongeant sa taille. (D) Après plusieurs événements de substitutions d'acides aminés et de combinaisons d'IGORFs via des mutations de codon STOP ou des insertions délétions, nous obtenons une protéine qui présente les caractéristiques des protéines canoniques (c'est-à-dire de longues séquences, de longs linkers, un enrichissement en résidus polaires et chargés) qui permettent l'optimisation de sa flexibilité, l'augmentation de la spécificité de son processus de repliement, de son repliement 3D et de ses interactions et enfin qui participent avec des événements de réarrangement ou de duplication de domaines dans l'architecture modulaire de protéines. Il convient de noter que bien que la figure se concentre sur l'émergence de gènes de novo, ce modèle peut également s'appliquer à des protéines déjà existantes.

Dans la partie suivante de ma thèse, en utilisant des approches de phylostratigraphy, nous avons divisé les protéines de *S. cerevisiae* en fonction de leur âge phylogénétique relatif afin d'étudier à quelle vitesse sont fixées et comment évoluent au cours de l'évolution, les propriétés de séquence et de structure des protéines que nous avons identifiées. Nous avons alors pu montrer que les jeunes protéines de *S. cerevisiae* présentent des propriétés intermédiaires entre les peptides potentiellement encodés par les IGORFs et les protéines plus anciennes. Notamment, bien que les protéines du protéome de *S. cerevisaie* deviennent continuellement plus longues au cours de l'évolution, les protéines plus anciennes présentent des tailles de linkers similaires tandis que les jeunes protéines codées présentent des tailles de linkers intermédiaires entre les IGORF et les protéines plus anciennes, soutenant ainsi que la taille des linkers est une propriété fixée au début de l'évolution des protéines. D'autre part, les clusters HCA de toutes les protéines de *S. cerevisiae* présentent des tailles statistiquement similaires, peu importe l'âge des protéines, renforçant ainsi le concept de clusters hydrophobes en tant que briques élémentaires de construction des protéines.

Dans la dernière partie de ma thèse, nous avons développé un modèle d'apprentissage supervisé qui vise à prédire le comportement de repliement (c'est-à-dire désordonné, stable en solution, stable lors de l'interaction avec un partenaire ou transmembranaire avec une forte propension à s'agréger en solution) des peptides potentiellement encodés par les IGORFs dans l'environnement cellulaire. L'objectif de cette partie était d'explorer plus finement les propriétés structurales des peptides codés par les IGORFs afin de mieux comprendre l'émergence des gènes de novo et d'étudier plus profondément comment l'expression omniprésente des IGORFs pourrait être tolérée par la cellule.

Nos résultats préliminaires montrent que les IGORFs codent pour des peptides potentiels avec une large gamme d'états de repliement, tandis que les peptides désordonnés ou propices à l'agrégation semblent être sous-représentés dans les peptides qui ont été montrés traduits par des approches expérimentales. De plus, nous avons identifié un nombre important de séquences d'acides aminés codées par les IGORFs qui n'est représenté par aucun état de repliement. Est-ce que ces séquences correspondent simplement à un biais technique dû à l'absence d'une catégorie de repliement dans notre ensemble d'apprentissage initial, ou est-ce qu'elles témoignent d'un réel état de repliement sous-représenté dans le monde codant ? Cette question est ouverte et motivera de futures analyses.

Au final, ma thèse m'a permis de développer des méthodes pour explorer les régions non-codantes avec un regard structural (étude des propriétés de structure des peptides que ces régions encodent potentiellement) mais aussi OMIQUE (étude de la traduction de ces régions). J'ai pu appliquer cet ensemble de méthodes à l'étude des régions non-codantes de *S. cerevisiae* et pu mettre en évidence que les IGORFs de la levure encodent des peptides présentant une grande variabilité de propriétés structurales incluant des peptides avec une forte propension au désordre, à l'agrégation ou avec un fort potentiel de repliement. Ensuite, j'ai pu mettre en évidence les déterminants de séquence et de structure pour l'émergence de nouveaux gènes. Ces travaux ouvrent la voie à de nombreuses études afin de tester la généralité de ces résultats sur d'autres espèces présentant des propriétés génomiques différentes (différents taux GC ou différentes compacités de génomes).

9 References

Abdi H, Williams LJ. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**: 433–459.

Abrusán G. 2013. Integration of New Genes into Cellular Networks, and Their Structural Maturation. *Genetics* **195**: 1407.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**: 403–410.

Alva V, Remmert M, Biegert A, Lupas AN, Söding J. 2010. A galaxy of folds. *Protein Science* **19**: 124–130.

Alva V, Söding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**: e09410.

Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, et al. 2015. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**: 595–606.

Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. 2014. SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research* **42**: D310–D314.

Argos P, Rao JM, Hargrave PA. 1982. Structural prediction of membrane-bound proteins. *European Journal of Biochemistry* **128**: 565–575.

Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso J-P. 2014. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *Elife* **3**: e03528.

Bae K, Mallick BK, Elsik CG. 2005. Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics* **21**: 2264–2270.

Barbosa C, Peixeiro I, Romão L. 2013. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet* **9**: e1003529.

Bartonek L, Braun D, Zagrovic B. 2020. Frameshifting preserves key physicochemical properties of proteins. *Proceedings of the National Academy of Sciences* **117**: 5907–5912.

Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. *PLoS computational biology* **13**: e1005375.

Basrai MA, Hieter P, Boeke JD. 1997. Small open reading frames: beautiful needles in the haystack. *Genome research* **7**: 768–771.

Bateson W, Waunders ER, Punnett RC. 1909. Experimental studies in the physiology of heredity. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **2**: 17–19.

Baudin-Baillieu A, Legendre R, Kuchly C, Hatin I, Demais S, Mestdagh C, Gautheret D, Namy O. 2014. Genome-wide translational changes induced by the prion [PSI+]. *Cell reports* **8**: 439–448.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal* **33**: 981–993.

Beadle GW, Tatum EL. 1941. Genetic control of biochemical reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* **27**: 499.

Belinky F, Babenko VN, Rogozin IB, Koonin EV. 2018. Purifying and positive selection in the evolution of stop codons. *Scientific reports* **8**: 1–11.

Berezovsky IN. 2019. Towards descriptor of elementary functions for protein design. *Current opinion in structural biology* **58**: 159–165.

Berezovsky IN, Grosberg AY, Trifonov EN. 2000. Closed loops of nearly standard size: common basic element of protein structure. *Febs Letters* **466**: 283–286.

Berezovsky IN, Kirzhner VM, Kirzhner A, Trifonov EN. 2001. Protein folding: looping from hydrophobic nuclei. *Proteins: Structure, Function, and Bioinformatics* **45**: 346–350.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic acids research* **28**: 235–242.

Bhaskaran R, Ponnuswamy P. 1988. Positional flexibilities of amino acid residues in globular proteins. *International Journal of Peptide and Protein Research* **32**: 241–255.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research* **42**: W252–W258.

Bitard-Feildel T, Callebaut I. 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Scientific reports* **7**: 1–13.

Bitard-Feildel T, Callebaut I. 2018a. HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences. *bioRxiv* 249995.

Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. 2015. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis." *Biochimie* **119**: 244–253.

Bitard-Feildel T, Lamiable A, Mornon J, Callebaut I. 2018. Order in disorder as observed by the "hydrophobic cluster analysis" of protein sequences. *Proteomics* **18**: 1800054.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**: 291–336.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature communications* **12**: 1–13.

Blevins WR, Tavella T, Moro SG, Blasco-Moreno B, Closa-Mosquera A, Díez J, Carey LB, Albà MM. 2019. Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Scientific reports* **9**: 1–11.

Blouin C, Butt D, Roger AJ. 2004. Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. *Protein science* **13**: 608–616.

Bornberg-Bauer E, Alba MM. 2013. Dynamics and adaptive benefits of modular protein evolution. *Current opinion in structural biology* **23**: 459–466.

Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally evolved de novo proteins. *Current Opinion in Structural Biology* **68**: 175–183.

Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from 'dark genomic matter'by 'grow slow and moult.' *Biochemical Society Transactions* **43**: 867–873.

Boveri T. 1903. Über die Konstitution der chromatischen Kernsubstanz. *Verh Dtsch Zool Ges* **13**: 10–33.

Boveri T. 1902. Uber mehrpolige mitosen als mittle zur analyse des zellkerns. *Verhandl Phys-med Ges (Wulzburg) NF* **35**: 67–90.

Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature reviews Molecular cell biology* **16**: 651–664.

Brennicke A, Marchfelder A, Binder S. 1999. RNA editing. *FEMS microbiology reviews* **23**: 297–316.

Bresler SE, Talmud D. 1944. On the nature of globular proteins. *CR Acad Sci USSR* **43**: 310–314.

Brooks DJ, Fresco JR. 2003. Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. *Gene* **303**: 177–185.

Brunet MA, Brunelle M, Lucier J-F, Delcourt V, Levesque M, Grenier F, Samandi S, Leblanc S, Aguilar J-D, Dufour P, et al. 2019. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Research* **47**: D403–D410.

Brunet MA, Lucier J-F, Levesque M, Leblanc S, Jacques J-F, Al-Saedi HRH, Guilloy N, Grenier F, Avino M, Fournier I, et al. 2021. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Research* **49**: D380–D388.

Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MH. 2017. Foldability of a natural de novo evolved protein. *Structure* **25**: 1687–1696.

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, et al. 2021. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research* **49**: D437–D451.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* **179**: 487–496.

Carlson EA. 1966. The gene: a critical history. *The gene: a critical history*.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *science* **309**: 1559–1563.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370–374.

Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potential: hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *Journal of molecular biology* **224**: 725–732.

Chandonia J-M, Fox NK, Brenner SE. 2019. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research* **47**: D475–D481.

Charton M. 1981. Protein folding and the genetic code: an alternative quantitative model. *Journal of theoretical biology* **91**: 115–123.

Charton M, Charton BI. 1983. The dependence of the Chou-Fasman parameters on amino acid side chain structure. *Journal of theoretical biology* **102**: 121–134.

Charton M, Charton BI. 1982. The structural dependence of amino acid hydrophobicity parameters. *Journal of theoretical biology* **99**: 629–644.

Cheek S, Krishna SS, Grishin NV. 2006. Structural classification of small, disulfide-rich protein domains. *Journal of molecular biology* **359**: 215–237.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**: 1140–1146.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40**: D700–D705.

Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. 2013. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* **140**: 2828–2834.

Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL. 2006. Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chemistry & biodiversity* **3**: 840–859.

Choi I-G, Kim S-H. 2006. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences* **103**: 14056–14061.

Chou, Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology - and Related Areas of Molecular Biology* **47**: 45–148.

Cid H, Bunster M, Canales M, Gazitúa F. 1992. Hydrophobicity and structural classes in proteins. *Protein Engineering, Design and Selection* **5**: 373–375.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. 2011. The reality of pervasive transcription. *PLoS Biol* **9**: e1000625.

Comings DE. 1972. The structure and function of chromatin. *Advances in human genetics* 237–431.

Cosic I. 1994. Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications. *IEEE Transactions on Biomedical Engineering* **41**: 1101–1114.

Couso J-P, Patraquim P. 2017. Classification and function of small open reading frames. *Nature reviews Molecular cell biology* **18**: 575–589.

Cuevas MVR, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, Lanoix J, Côté C, Staudt LM, Lemieux S. 2021. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell reports* **34**: 108815.

Davidson AR, Lumb KJ, Sauer RT. 1995. Cooperatively folded proteins in random sequence libraries. *Nature structural biology* **2**: 856–864.

Davis CA, Ares M. 2006. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomycescerevisiae. *Proceedings of the National Academy of Sciences* **103**: 3262–3267.

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. 2017. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research* **45**: D289–D295.

Deng Y, Bamigbade AT, Hammad MA, Xu S, Liu P. 2018. Identification of small ORF-encoded peptides in mouse serum. *Biophysics reports* **4**: 39–49.

Dosztányi Z. 2018. Prediction of protein disorder based on IUPred. *Protein Science* **27**: 331–340.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology* **347**: 827–839.

Durand É, Gagnon-Arsenault I, Hallin J, Hatin I, Dubé AK, Nielly-Thibault L, Namy O, Landry CR. 2019. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome research* **29**: 932–943.

Edwards H, Abeln S, Deane CM. 2013. Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS computational biology* **9**: e1003325.

Eguen T, Straub D, Graeff M, Wenkel S. 2015. MicroProteins: small size–big impact. *Trends in plant science* **20**: 477–482.

Ekman D, Elofsson A. 2010. Identifying and quantifying orphan protein sequences in fungi. *Journal of molecular biology* **396**: 396–405.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature* **447**: 799.

Erdős G, Dosztányi Z. 2020. Analyzing Protein Disorder with IUPred2A. *Current Protocols in Bioinformatics* **70**: e99.

Espadaler J, Querol E, Aviles FX, Oliva B. 2006. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics* **22**: 2237–2243.

Fasman GD. 1975. Handbook of biochemistry and Molecular Biology. *Nucleic Acids* **1**: 589.

Fauchere J-L, Charton M, Kier LB, Verloop A, Pliska V. 1988. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research* **32**: 269–278.

Faure G, Callebaut I. 2013a. Comprehensive repertoire of foldable regions within whole genomes. *PLoS computational biology* **9**.

Faure G, Callebaut I. 2013b. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics* **29**: 1726–1733.

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology* **22**: 1302–1306.

Ferruz N, Noske J, Höcker B. 2021. Protlego: a Python package for the analysis and design of chimeric proteins. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab253 (Accessed August 4, 2021).

Fiers W, Contreras R, De Wachter R, Haegeman G, Merregaert J, Jou WM, Vandenberghe A. 1971. Recent progress in the sequence determination of bacteriophage MS2 RNA. *Biochimie* **53**: 495–506.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Jou WM, Molemans F, Raeymaekers A, Van den Berghe A, et al. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**: 500–507.

Fiser A, Šali A. 2003. Modeller: generation and refinement of homology-based protein structure models. *Methods in enzymology* **374**: 461–491.

Flemming W. 1882. *Zellsubstanz, kern und zelltheilung*. Vogel.

Foy SG, Wilson BA, Bertram J, Cordes MH, Masel J. 2019. A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. *Genetics* **211**: 1345–1355.

Fozo EM, Hemm MR, Storz G. 2008. Small toxic proteins and the antisense RNAs that repress them. *Microbiology and Molecular Biology Reviews* **72**: 579–589.

François P, Arbes H, Demais S, Baudin-Baillieu A, Namy O. 2021. RiboDoc: A Docker-based package for ribosome profiling analysis. *Computational and Structural Biotechnology Journal* **19**: 2851–2860.

Ganesan A, Siekierska A, Beerten J, Brams M, Van Durme J, De Baets G, Van der Kant R, Gallardo R, Ramakers M, Langenberg T. 2016. Structural hot spots for the solubility of globular proteins. *Nature communications* **7**: 1–15.

Gardner LB. 2008. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Molecular and cellular biology* **28**: 3729–3741.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome research* **17**: 669–681.

Ghouzam Y, Postic G, Guerin P-E, de Brevern AG, Gelly J-C. 2016. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Scientific Reports* **6**: 1–10.

GitHub. 2021. *GitHub*. https://github.com/.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of molecular biology* **313**: 903–919.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**: 862–864.

Greenwald J, Riek R. 2012. On the possible amyloid origin of protein folds. *Journal of molecular biology* **421**: 417–426.

Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. 2017. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. *Molecular biology and evolution* **34**: 1066–1082.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**: 307–321.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology* **28**: 503–510.

Hanada K, Zhang X, Borevitz JO, Li W-H, Shiu S-H. 2007. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome research* **17**: 632–640.

Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, Bassot C, Benítez GI, Bevilacqua M, Chasapi A, et al. 2020. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Research* **48**: D269–D276.

Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**: 1323–1330.

Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. *Journal of molecular evolution* **88**: 382–398.

Heinen TJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Current Biology* **19**: 1527–1531.

Heinig M, Frishman D. 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* **32**: W500–W502.

Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular microbiology* **70**: 1487–1501.

Hobbs EC, Fontaine F, Yin X, Storz G. 2011. An expanding universe of small proteins. *Current opinion in microbiology* **14**: 167–173.

Höcker B. 2014. Design of proteins from smaller fragments—learning from evolution. *Current opinion in structural biology* **27**: 56–62.

Hsu PY, Calviello L, Wu H-YL, Li F-W, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proceedings of the National Academy of Sciences* **113**: E7126–E7135.

Hu B, Xie G, Lo C-C, Starkenburg SR, Chain PS. 2011. Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Briefings in functional genomics* **10**: 322–333.

Huang X, Miller W. 1991. A time-efficient, linear-space local similarity algorithm. *Advances in applied mathematics* **12**: 337–357.

Hubbard TJ, Murzin AG, Brenner SE, Chothia C. 1997. SCOP: a structural classification of proteins database. *Nucleic acids research* **25**: 236–239.

Ibstedt S, Sideri TC, Grant CM, Tamás MJ. 2014. Global analysis of protein aggregation in yeast during physiological conditions and arsenite stress. *Biology open* **3**: 913–923.

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. 2014. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* **8**: 1365–1379.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science* **324**: 218–223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789–802.

Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, Khan OM, Brewer JR, Skadow MH, Duizer C, et al. 2018. The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**: 434–438.

Jacob F. 1977. Evolution and tinkering. *Science* **196**: 1161–1166.

Janin J, Wodak S, Levitt M, Maigret B. 1978. Conformation of amino acid side-chains in proteins. *Journal of molecular biology* **125**: 357–386.

Jensen TH, Jacquier A, Libri D. 2013. Dealing with pervasive transcription. *Molecular cell* **52**: 473–484.

Johannsen W. 1909. *Elemente der exakten Erblichkeitslehre*. Fischer.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**: 275–282.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.

Jungck JR. 1978. The genetic code as a periodic table. *Journal of Molecular Evolution* **11**: 211–224.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome research* **20**: 1313–1326.

Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**: 1096–1099.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916–919.

Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature* **410**: 715–718.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* **111**: 6131–6138.

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. 2018. Origins and structural properties of novel and de novo protein domains during insect evolution. *The FEBS journal* **285**: 2605–2625.

Klein P, Kanehisa M, DeLisi C. 1984. Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* **787**: 221–226.

Knopp M, Gudmundsdottir JS, Nilsson T, König F, Warsi O, Rajer F, Ädelroth P, Andersson DI. 2019. De novo emergence of peptides that confer antibiotic resistance. *MBio* **10**: e00837-19.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome research* **19**: 1752–1759.

Koehl P, Levitt M. 1999. Structure-based conformational preferences of amino acids. *Proceedings of the National Academy of Sciences* **96**: 12524–12529.

Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. 2021. Bridging themes: short protein segments found in different architectures. *Molecular biology and evolution* **38**: 2191–2208.

Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *Journal of Biological Chemistry* **289**: 30334–30342.

Kozma D, Simon I, Tusnady GE. 2012. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic acids research* **41**: D524–D529.

Krigbaum W, Komoriya A. 1979. Local interactions as a structure determinant for protein molecules: II. *Biochimica et biophysica acta* **576**: 204–248.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**: 567–580.

Kuhlman B, Bradley P. 2019. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **20**: 681–697.

Kurosaki T, Popp MW, Maquat LE. 2019. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nature reviews Molecular cell biology* **20**: 406–420.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**: 105–132.

LaBean TH, Butt TR, Kauffman SA, Schultes EA. 2011. Protein folding absent selection. *Genes* **2**: 608–626.

Lamarine M, Mornon J-P, Berezovsky IN, Chomilier J. 2001. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cellular and Molecular Life Sciences CMLS* **58**: 492–498.

Lamiable A, Bitard-Feildel T, Rebehmed J, Quintus F, Schoentgen F, Mornon J-P, Callebaut I. 2019. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* **167**: 68–80.

Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. 2021. Structural and functional characterization of a putative de novo gene in Drosophila. *Nature communications* **12**: 1–13.

Langenberg T, Gallardo R, van der Kant R, Louros N, Michiels E, Duran-Romaña R, Houben B, Cassio R, Wilkinson H, Garcia T, et al. 2020. Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins. *Cell reports* **31**: 107512.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: 1–10.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214–218.

Leff SE, Rosenfeld MG, Evans RM. 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. *Annual review of biochemistry* **55**: 1091–1117.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences* **103**: 9935–9939.

Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of molecular biology* **104**: 59–107.

Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences* **109**: 20461–20466.

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell research* **20**: 408–420.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li J, Liu C. 2019. Coding or noncoding, the converging concepts of RNAs. *Frontiers in Genetics* **10**: 496.

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009b. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal* **58**: 485–498.

Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the origin of de novo genes in Arabidopsis thaliana populations. *Genome biology and evolution* **8**: 2190–2202.

Lin M, Lucas Jr HC, Shmueli G. 2013. Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research* **24**: 906–917.

Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *Journal of molecular biology* **342**: 345–353.

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC bioinformatics* **11**: 579.

Lu T-C, Leu J-Y, Lin W-C. 2017. A comprehensive analysis of transcript-supported de novo genes in Saccharomyces sensu stricto yeasts. *Molecular biology and evolution* **34**: 2823–2838.

Lumb KJ, Kim PS. 1995. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**: 8642–8648.

Lybecker M, Bilusic I, Raghavan R. 2014. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription* **5**: e944039.

Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M, Saghatelian A. 2014. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *Journal of proteome research* **13**: 1757–1765.

Macossay-Castillo M, Marvelli G, Guharoy M, Jain A, Kihara D, Tompa P, Wodak SJ. 2019. The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity. *Journal of molecular biology* **431**: 1650–1670.

Magny Emile G., Pueyo Jose Ignacio, Pearl Frances M.G., Cespedes Miguel Angel, Niven Jeremy E., Bishop Sarah A., Couso Juan Pablo. 2013. Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* **341**: 1116–1120.

Makarewich CA, Olson EN. 2017. Mining for micropeptides. *Trends in cell biology* **27**: 685–696.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17**: 10–12.

Maupetit J, Derreumaux P, Tuffery P. 2009. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research* **37**: W498–W503.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**: 20140332.

Mendel G. 1865. Versuche uber Pflanzen-hybriden. Verh. Naturf. Ver. *Brunn, Bd* **10**: 1865.

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research* **46**: W329–W337.

Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* **5**: e1000376.

Mezei M. 2020. On predicting foldability of a protein from its sequence. *Proteins: Structure, Function, and Bioinformatics* **88**: 355–365.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**: D412–D419.

Mitaku S, Hirokawa T, Tsuji T. 2002. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces. *Bioinformatics* **18**: 608–616.

Monné M, Nilsson I, Elofsson A, von Heijne G. 1999. Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *Journal of molecular biology* **293**: 807–814.

Muller HJ. 1927. Artificial transmutation of the gene. *Science* **66**: 84–87.

Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. *PloS one* **7**: e48650.

Namy O, Duchateau-Nguyen G, Hatin I, Hermann-Le Denmat S, Termier M, Rousset J. 2003. Identification of stop codon readthrough genes in Saccharomyces cerevisiae. *Nucleic acids research* **31**: 2289–2296.

Ndah E, Jonckheere V, Giess A, Valen E, Menschaert G, Van Damme P. 2017. REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Research* **45**: e168–e168.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**: 1038–1042.

Nelson B, Makarewich C, Anderson D, Winders B, Troupes C, Fenfen W, Reese A, McAnally J, Xiongwen C, Kavalari E, et al. 2016. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**: 271–275.

Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant source of bioactive RNAs or peptides. *Nature ecology & evolution* **1**: 1–7.

Nepomnyachiy S, Ben-Tal N, Kolodny R. 2017. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings of the National Academy of Sciences* **114**: 11703–11708.

Nielly-Thibault L, Landry CR. 2019. Differences between the raw material and the products of de novo gene birth can result from mutational biases. *Genetics* **212**: 1353–1366.

Nishikawa K, Ooi T. 1980. Prediction of the surface-interior diagram of globular proteins by an empirical method. *International journal of peptide and protein research* **16**: 19–32.

Nishikawa K, Ooi T. 1986. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *The Journal of Biochemistry* **100**: 1043–1047.

Ohno S. 2013. *Evolution by gene duplication*. Springer Science & Business Media.

Ohno S. 1972. So much'junk'DNA in our genome. pp. 366–370.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563–573.

Orr MW, Mao Y, Storz G, Qian S-B. 2020. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research* **48**: 1029–1042.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. *elife* **3**: e01311.

Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. 2021. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Research* **31**: 2303–2315.

Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. 2016. The role of protein loops and linkers in conformational dynamics and allostery. *Chemical reviews* **116**: 6391–6423.

Papandreou N, Berezovsky IN, Lopes A, Eliopoulos E, Chomilier J. 2004. Universal positions in globular proteins: From observation to simulation. *European journal of biochemistry* **271**: 4762–4768.

Peng J, Xu J. 2011. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics* **79**: 161–171.

Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, et al. 2017. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic acids research* **45**: D219–D227.

Ponnuswamy P. 1993. Hydrophobic characteristics of folded proteins. *Progress in biophysics and molecular biology* **59**: 57–103.

Ponnuswamy P, Prabhakaran M, Manavalan P. 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **623**: 301–316.

Portin P, Wilkins A. 2017. The evolving definition of the term "gene." *Genetics* **205**: 1353–1364.

Postic G, Ghouzam Y, Chebrek R, Gelly J-C. 2017. An ambiguity principle for assigning protein structural domains. *Science advances* **3**: e1600552.

Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. 2012. Stop codons in bacteria are not selectively equivalent. *Biology direct* **7**: 1–13.

Prabakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, Dittrich C, Hong E, Gunawardena J, Steen H, Kreiman G, et al. 2014. Quantitative profiling of peptides from RNAs classified as noncoding. *Nature communications* **5**: 1–10.

Prabh N, Rödelsperger C. 2019. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in pristionchus nematodes. *G3: Genes, Genomes, Genetics* **9**: 2277–2286.

Rackovsky S, Scheraga H. 1977. Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. *Proceedings of the National Academy of Sciences* **74**: 5248–5251.

Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* **167**: 122–132.

Radzicka A, Wolfenden R. 1988. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**: 1664–1670.

Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular biology and evolution* **35**: 2582–2584.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS one* **6**: e22594.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature biotechnology* **29**: 24–26.

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**: 834–838.

Roseman MA. 1988. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *Journal of molecular biology* **200**: 513–522.

Rousseau F, Schymkowitz J, Serrano L. 2006a. Protein aggregation and amyloidosis: confusion of the kinds? *Current opinion in structural biology* **16**: 118–126.

Rousseau F, Serrano L, Schymkowitz JW. 2006b. How evolutionary pressure against protein aggregation shaped chaperone specificity. *Journal of molecular biology* **355**: 1037–1047.

Ruiz-Orera J, Albà MM. 2019a. Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures. *NAR genomics and bioinformatics* **1**: e2–e2.

Ruiz-Orera J, Albà MM. 2019b. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends in Genetics* **35**: 186–198.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas J, Messeguer X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat. Ecol. Evol. 2: 890–896. *Nature ecology & evolution*.

Samayoa J, Yildiz FH, Karplus K. 2011. Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics* **27**: 1765–1771.

Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. 2019. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**: 510–514.

Scannell D, Zill O, Rokas A, Payen C, Dunham M, Eisen M, Rine J, Johnston M, Hittinger C, et al. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 (Bethesda). 2011; 1 (1): 11–25. *Genetics Society of America*.

Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. 2018. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**: 535–537.

Schaefer C, Schlessinger A, Rost B. 2010. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics* **26**: 625–631.

Schavemaker PE, Śmigiel WM, Poolman B. 2017. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. *Elife* **6**: e30084.

Schlötterer C. 2015. Genes from scratch–the evolutionary fate of de novo genes. *Trends in Genetics* **31**: 215–219.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature ecology & evolution* **2**: 1626–1632.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic acids research* **33**: W382–W388.

Sendoel A, Dunn JG, Rodriguez EH, Naik S, Gomez NC, Hurwitz B, Levorse J, Dill BD, Schramek D, Molina H, et al. 2017. Translation from unconventional 5′ start sites drives tumour initiation. *Nature* **541**: 494–499.

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**: 706–710.

Siepel A. 2009. Darwinian alchemy: Human genes from noncoding DNA. *Genome research* **19**: 1693–1695.

Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CS, Woodridge L, Rauer C, Sen N, et al. 2021. CATH: increased structural coverage of functional space. *Nucleic acids research* **49**: D266–D273.

Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, Tuknait A, Agrawal P, Mathur D, Raghava GPS. 2016. SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Research* **44**: D1119–D1126.

Skrzypek MS, Hirschman J. 2011. Using the Saccharomyces Genome Database (SGD) for analysis of genomic information. *Current protocols in bioinformatics* **35**: 1–20.

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. 2013. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology* **9**: 59.

Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, Coller J, Baker KE. 2014. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell reports* **7**: 1858–1866.

Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**: W244–W248.

Sonnhammer EL, Von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Vol. 6 of, pp. 175–182.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**: D535–D539.

Storz G, Wolf YI, Ramamurthi KS. 2014. Small proteins can no longer be ignored. *Annual review of biochemistry* **83**: 753–777.

Sutton W. 1903. The Chromosomes in Heredity. *Biological Bulletin* **4**: 231–251.

Takano K, Yutani K. 2001. A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Engineering* **14**: 525–528.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**: 692–702.

Team R Core RC. 2020. R: A language and environment for statistical computing. https://www.R-project.org/.

Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP. 2004. Clustering of protein structural fragments reveals modular building block approach of nature. *Journal of molecular biology* **338**: 611–629.

Thiaville PC, Legendre R, Rojas-Benítez D, Baudin-Baillieu A, Hatin I, Chalancon G, Glavic A, Namy O, de Crécy-Lagard V. 2016. Global translational impacts of the loss of the tRNA modification t6A in yeast. *Microbial cell* **3**: 29.

Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J, et al. 2017. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific reports* **7**: 1–9.

Tretyachenko V, Vymetal J, Neuwirthová T, Vondrasek J, Fujishima K, Hlouchova K. 2021. Structured proteins are abundant in unevolved sequence space. *bioRxiv*.

Trifonov E. 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *Journal of molecular biology* **194**: 643–652.

Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN. 2001. Distinct stages of protein evolution as suggested by protein sequence analysis. *Journal of molecular evolution* **53**: 394–401.

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* **596**: 590–596.

Tusnády GE, Dosztányi Z, Simon I. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic acids research* **33**: D275–D278.

Tusnády GE, Dosztányi Z, Simon I. 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **20**: 2964–2972.

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**: D506–D515.

Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines CP, Iannotta J, et al. 2020a. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications* **11**: 1–18.

Vakirlis N, Carvunis A-R, McLysaght A. 2020b. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Molecular biology and evolution* **35**: 631–645.

Vakirlis N, McLysaght A. 2019. Computational prediction of de novo emerged protein-coding genes. In *Computational Methods in Protein Evolution*, pp. 63–81, Springer.

Van Oss SB, Carvunis A-R. 2019. De novo gene birth. *PLoS genetics* **15**.

van Westen GJ, Swier RF, Cortes-Ciriano I, Wegner JK, Overington JP, IJzerman AP, van Vlijmen HW, Bender A. 2013a. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *Journal of cheminformatics* **5**: 1–20.

van Westen GJ, Swier RF, Wegner JK, IJzerman AP, van Vlijmen HW, Bender A. 2013b. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *Journal of cheminformatics* **5**: 1–11.

Verbeke F, De Craemer S, Debunne N, Janssens Y, Wynendaele E, Van de Wiele C, De Spiegeleer B. 2017. Peptides as quorum sensing molecules: measurement techniques and obtained levels in vitro and in vivo. *Frontiers in neuroscience* **11**: 183.

Vihinen M, Torkkila E, Riikonen P. 1994. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics* **19**: 141–149.

von Bohlen AE, Böhm J, Pop R, Johnson DS, Tolmie J, Stücker R, Morris-Rosendahl D, Scherer G. 2017. A mutation creating an upstream initiation codon in the SOX 9 5′ UTR causes acampomelic campomelic dysplasia. *Molecular genetics & genomic medicine* **5**: 261–268.

Waldeyer W. 1888. Über Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen. *Archiv für mikroskopische Anatomie* **32**: 1–122.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & cellular proteomics* **11**: 492–500.

Wang S, Mao C, Liu S. 2019. Peptides encoded by noncoding genes: challenges and perspectives. *Signal Transduction and Targeted Therapy* **4**: 1–12.

Warme PK, Morgan RS. 1978. A survey of amino acid side-chain interactions in 21 proteins. *Journal of molecular biology* **118**: 289–304.

Watson JD, Crick FH. 1953a. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.

Watson JD, Crick FH. 1953b. The structure of DNA. Vol. 18 of, pp. 123–131, Cold Spring Harbor Laboratory Press.

Weaver Jeremy, Mohammad Fuad, Buskirk Allen R., Storz Gisela, Vogel Joerg. 2019. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *mBio* **10**: e02819-18.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature ecology & evolution* **1**: 1–6.

Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome biology and evolution* **3**: 1245–1252.

Wissler L, Godmann L, Bornberg-Bauer E. 2012. Evolutionary dynamics of simple sequence repeats across long evolutionary time scale in genus Drosophila. *Trends in Evolutionary Biology* **4**: e7–e7.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences* **106**: 7273–7280.

Wolfenden R, Andersson L, Cullis P, Southgate C. 1981. Affinities of amino acid side chains for solvent water. *Biochemistry* **20**: 849–855.

Wu B, Knudson A. 2018. Tracing the de novo origin of protein-coding genes in yeast. *MBio* **9**.

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. *PLoS genetics* **7**: e1002379.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.

Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, Franci C, Cheung TK, Fritsche J, Weinschenk T, et al. 2014. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**: 572–576.

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**: 1496–1503.

Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, Adams RM, Shah MB, Hettich RL, Lindquist E, et al. 2011. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome research* **21**: 634–641.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution* **17**: 32–43.

Yao R-W, Wang Y, Chen L-L. 2019. Cellular functions of long noncoding RNAs. *Nature cell biology* **21**: 542–551.

Yin M, Goncearenco A, Berezovsky IN. 2021. Deriving and using descriptors of elementary functions in rational protein design. *Frontiers in Bioinformatics* **1**: 8.

Yin X, Jing Y, Xu H. 2019. Mining for missed sORF-encoded peptides. *Expert Review of Proteomics* **16**: 257–266.

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature genetics* **49**: 913–924.

Zetoune AB, Fontanière S, Magnin D, Anczuków O, Buisson M, Zhang CX, Mazoyer S. 2008. Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC genetics* **9**: 1–11.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019a. Rapid evolution of protein diversity by de novo origination in Oryza. *Nature ecology & evolution* **3**: 679–690.

Zhang W, Gao Y, Long M, Shen B. 2019b. Origination and evolution of orphan genes and de novo genes in the genome of Caenorhabditis elegans. *Science China Life Sciences* **62**: 579–593.

Zhang Z, Lu M, Qin Y, Gao W, Tao L, Su W, Zhong J. 2021. Neoantigen: A New Breakthrough in Tumor Immunotherapy. *Frontiers in Immunology* **12**: 1297.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343**: 769–772.

Zhou H, Zhou Y. 2004. Quantifying the effect of burial of amino acid residues on protein stability. *PROTEINS: Structure, Function, and Bioinformatics* **54**: 315–322.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome research* **18**: 1446–1455.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Le génome non codant, réservoir de nouveauté génétique

Mots clés : gènes de novo, génome non codant, potentiel de repliement, évolution des protéines, briques protéiques

Résumé : Le génome non codant joue un rôle important dans la naissance des gènes de novo et l'émergence de la nouveauté génétique. Tous les gènes codant pour des protéines, quelle que soit leur histoire évolutive, doivent avoir eu à un moment donné une forme ancestrale non-codante. D'autre part, les protéomes sont caractérisés par une grande diversité d'états structuraux. Néanmoins, la façon dont les propriétés des séquences non-codantes permettent la naissance de nouveaux gènes et façonnent la diversité structurale et l'évolution des protéines demeure inconnu.

Au cours de ma thèse, combinant différentes approches de bioinformatiques, j'ai caractérisé la diversité du potentiel de repliement des séquences d'acides aminés encodées par tous les ORF (Open Reading Frames) intergéniques de S. cerevisiae dans le but (i) d'explorer si la diversité des états structuraux des protéomes est aussi retrouvée dans les ORFs non-codantes, et (ii) d'estimer le potentiel du génome non-codant à produire de nouvelles briques protéiques qui peuvent soit donner naissance à de nouveaux gènes, soit être intégrées dans des protéines préexistantes, participant ainsi à la diversité et à l'évolution des structures protéiques.
J'ai montré que les séquences d'acides aminés encodées par la plupart des ORF intergéniques contiennent les blocs élémentaires des structures protéiques. Ces derniers correspondent à des groupements de résidus riches en acides aminés hydrophobes. De plus, j'ai montré que ces séquences couvrent la grande diversité d'états structuraux des protéines canoniques, avec la majorité d'entre elles prédites comme repliables.

De plus, en utilisant des approches de reconstruction ancestrale ainsi que des expériences de « ribosome footprint profiling », j'ai identifié des caractéristiques de séquence et de structure déterminant l'émergence de nouveaux gènes.

En particulier, j'ai montré une forte corrélation entre le potentiel de repliement des protéines de novo et celui de leurs séquences ancestrales, reflétant ainsi la relation entre le génome non-codant et l'univers des structures protéiqes. L'ensemble de ces résultats m'a permis de proposer un modèle de naissance de genes de novo et d'évolution de protéines à partir de régions non-codantes reposant sur les ORF intergéniques comme modules structuraux élémentaires.

De plus, en utilisant des approches de phylostragraphie, j'ai pu classer les protéines de la levure en différents phylostrates en fonction de leur âge évolutif relatif. Cela m'a permis d'étudier à quelle vitesse se fixent et par conséquent comment évoluent diverses propriétés de séquence et de structure des protéines de la levure au cours de l'évolution.

Dans la dernière partie de ma thèse, je me suis intéressé à mieux caractériser l'état de repliement des peptides potentiellement codés par les petits ORF d'un génome non-codant (par exemple, désordonné, stable en solution, stable en interaction avec un partenaire ou transmembranaire) et donc prédire leur potentiel « comportement » dans l'environnement cellulaire. Pour ce faire, j'ai développé une méthode rapide basée sur des approches d'apprentissage automatique qui permet d'associer une courte séquence d'acides aminés à l'un des états de repliement prédéfinis. La méthode présente de très bonnes performances sur des peptides pour lesquels nous avons une caractérisation expérimentale. Je l'ai ensuite appliquée sur l'ensemble des ORF intergéniques de la levure et j'ai ainsi pu annoter la majorité d'entre eux avec l'une des catégories prédéfinies et j'ai pu confirmer fait qu'ils présentaient une grande diversité d'états structuraux.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : The noncoding genome, a reservoir of genetic novelty

Keywords : de novo genes, noncoding genome, foldability, protein evolution, protein bricks

Abstract : The noncoding genome plays an important role in de novo gene birth and the emergence of genetic novelty. All protein-coding genes, whatever their evolutionary history, must have had at some point a noncoding ancestral form. On the other hand, proteomes are characterized by a large diversity of structural states. Nevertheless, how the properties of noncoding sequences could promote the birth of novel genes and shape the structural diversity and evolution of proteins remains unclear.

During my thesis, combining different bioinformatic approaches, I characterized the fold potential diversity of the amino acid sequences encoded by all intergenic ORFs (Open Reading Frames) of S. cerevisiae with the aim of (i) exploring whether the structural states' diversity of proteomes is already present in noncoding sequences, and (ii) estimating the potential of the noncoding genome to produce novel protein bricks that can either give rise to novel genes or be integrated into pre-existing proteins, thus participating in protein structure diversity and evolution.
I found that the amino acid sequences encoded by most intergenic ORFs contain the elementary building blocks of protein structures corresponding to clusters rich in hydrophobic amino acids. Moreover, I showed that they encompass the large structural state diversity of canonical proteins with strikingly the majority of them predicted as foldable.

Furthermore, using ancestral reconstruction approaches together with ribosome footprint profiling experiments I identified sequence and structural features that determine the emergence of novel genes.
In particular, I observed a strong correlation between the fold potential of de novo proteins and the one of their ancestral amino acid sequences, reflecting the relationship between the noncoding genome and the protein structure universe. All these results permitted me to propose a model of de novo genes birth and protein evolution from noncoding regions with intergenic ORFs as elementary structural modules.

In addition, using phylostragraphy approaches, I was able to classify the yeast proteins into different phylostrata based on their relative evolutionary age. This permitted me to study how fast are fixed and consequently how evolve various sequence and structural properties of the yeast proteins along the evolutionary time.

In the final part of my thesis, I was interested at better characterizing the fold state of peptides potentially encoded by the small ORFs of a noncoding genome (i.e., disordered, stable in solution, stable upon interaction with a partner or transmembrane) and therefore predict their potential "behavior" in the cellular environment. To do so, I developed a fast method based on machine learning approaches which enables the association of a short amino acid sequence with one of the predefined fold states. The method presented very good performance on peptides with experimental characterization. I, then, applied it on the whole set of intergenic ORFs of the yeast permitting the annotation of the majority of them with one of the predefined fold state categories and confirming their vast structural state diversity.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France