Chapter 3

Exploring the Peptide Potential of Genomes

Chris Papadopoulos, Nicolas Chevrollier, and Anne Lopes

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 (1) their impact on the cell and (2) 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 produce 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 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 that of the nontranslated lncRNAs and more generally from the other noncoding ORFs of the mouse.

Key words Noncoding DNA, Fold potential, De novo genes, Small ORF-encoded peptides, ORF-track, 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 has been shown to be translated by ribosome profiling experiments [1–4]. In addition, mass spectrometry experiments conducted on mammals, bacteria, or plants [5–11] confirm the existence of these translation products in the cell,

Thomas Simonson (ed.), *Computational Peptide Science: Methods and Protocols*,
Methods in Molecular Biology, vol. 2405, https://doi.org/10.1007/978-1-0716-1855-4_3,

with the identification of hundreds of peptides derived from noncoding regions. The fact that these noncanonical products exhibit 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 [5, 6] 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 (1) be tolerated by the cell and (2) 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 [5, 12]. 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 them being even beneficial in *Escherichia coli* [13–16]. All these results place the foldability of noncoding ORFs at the center of novel gene birth and strengthen the need to characterize 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 (1) the extraction and annotation of all possible ORFs of a genome and (2) 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 [17–20], Tango [21–23], and IUPred2A [24–26], 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 that are able to fold to a stable

3D structure or to a molten globule state, in which the specific tertiary structure is lost but 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 to the small genome of *E. 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 annotation of all the possible ORFs present in noncoding regions. The ORFmine package is not published yet, but is available at: https://github.com/i2bc/ORFmine. It consists of two independent programs, ORFtrack and ORFold, that can be combined together or used independently (Fig. 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.

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

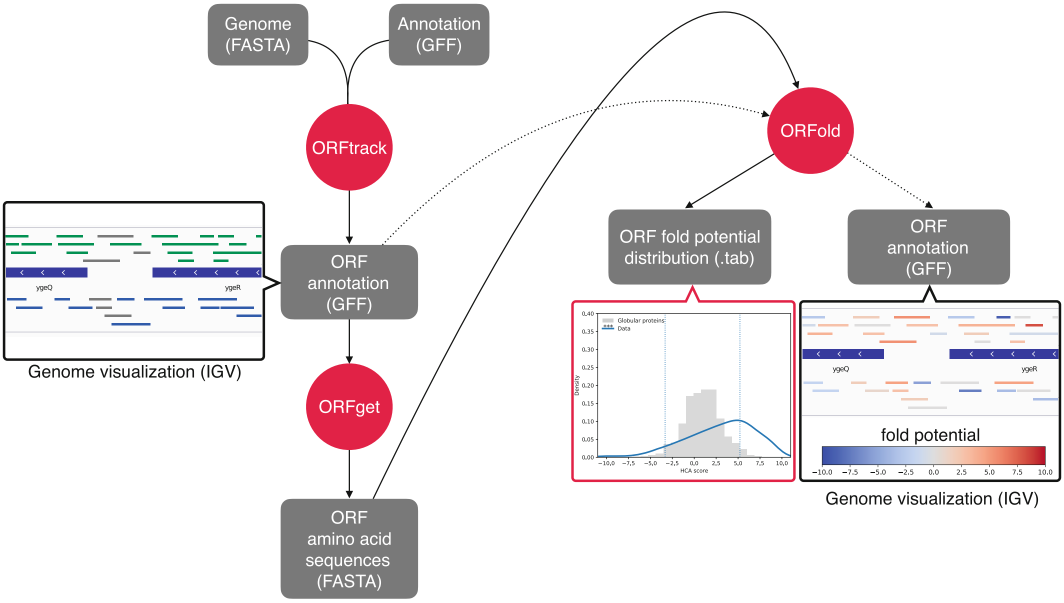

Fig. 1 Pipeline of ORFmine. The inputs and outputs are represented with gray rectangles while the main scripts are shown with red circles. The mandatory inputs necessary to the ORF annotation and the estimation of their structural properties (e.g., fold potential and disorder and aggregation propensities), 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 to 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)

resulting ORF (e.g., intergenic ORF, noncoding ORF that overlaps a coding sequence, coding ORF), 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) (Figs. 2 and 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 as a CDS). They are generally larger than the CDS since they are defined from 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) (Figs. 2 and 3). 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

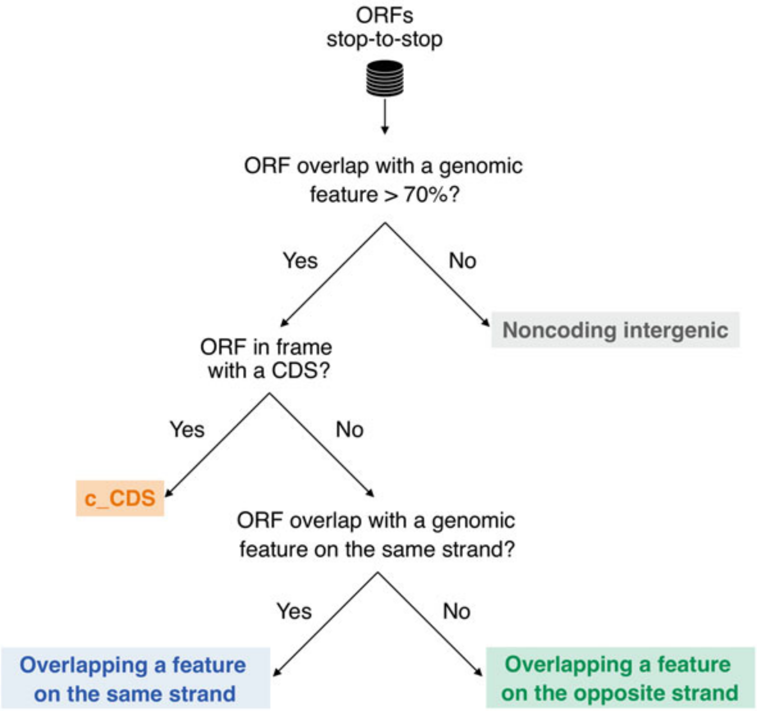
ORFs
stop-to-stop
ORF overlap with a genomic feature > 70%?
Yes
No
Noncoding intergenic
ORF in frame with a CDS?
Yes
No
c_CDS
ORF overlap with a genomic feature on the same strand?
Yes
No
Overlapping a feature on the same strand
Overlapping a feature on the opposite strand

Fig. 2 Decision tree of ORFtrack. ORFs are annotated according to four main categories: c_CDS for coding ORFs (orange box), noncoding intergenic ORFs (gray box), and noncoding ORFs that overlap a genomic feature on the same strand (blue box) or on the opposite strand (green box)

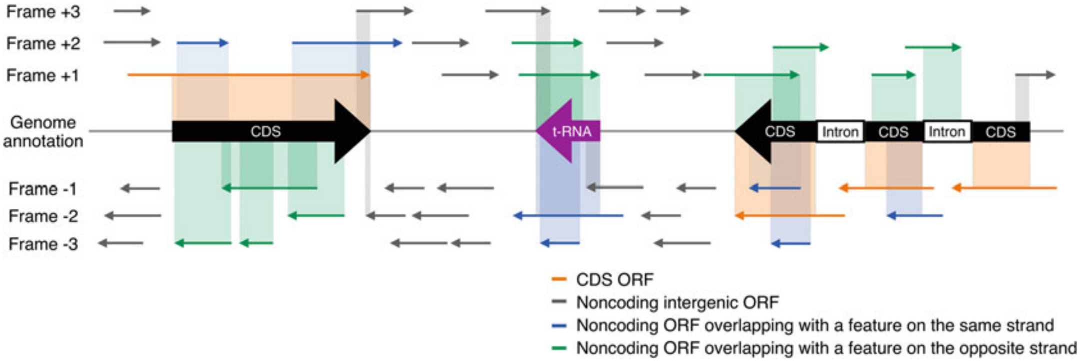

Fig. 3 Schematic representation of the six frames of a DNA section. The genomic features annotated in the original GFF file are represented in the middle line. The ORFs of the six frames are colored with respect to their ORFtrack annotation. The overlap between an ORF and a genomic feature is illustrated with a rectangle colored according to the ORF annotation

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:

1. 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).
2. 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).
3. 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).

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. An example is nc_ovp_same-lncRNAs and nc_ovp_opp-lncRNAs, if 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 [17–20]. In addition, it can predict their disorder or aggregation propensities, with IUPred and Tango, respectively [21–26]. 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 h on a single CPU (2 GHz processor, 16 GB RAM) for all the ORFs of *E. 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 to the GFF file. The latter can be uploaded subsequently on a genome viewer such as IGV [27], 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. [24].

HCA

ORFold estimates the fold potential with the HCA (Hydrophobic Cluster Analysis) approach [19, 28]. HCA toolkit is available at https://github.com/T-B-F/pyHCA. It splits an amino acid sequence into hydrophobic clusters and linkers. The former gathers strong hydrophobic residues (V, I, L, F, M, Y, W) and cysteines while the latter corresponds 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 by 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 Fig. 5.

Tango

ORFold calculates the aggregation propensity of a sequence with Tango [21–23], which is available at http://tango.crg.es upon request to the developers. Following the criteria proposed by Linding et al. [21], 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 [24–26, 29]. We use the version 2A of IUPred [24, 25], which is available at https://iupred2a.elte.hu upon request to the developers. Consistent 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 *E. coli* str. K-12 substr. MG1655 (*E. coli*), regardless whether 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

1. E_coli.fna (available at https://github.com/i2bc/ORFmine in the “examples” directory).
2. 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 ORFtrack instruction displays all the genomic features annotated in the *E. coli* genome:

```
> 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 (2 GHz processor, 16 GB RAM) is 38 s. ORFtrack generates a new GFF file (mapping_orf_E_coli.gff) that contains 135097 annotated ORFs of which 130637 are annotated as noncoding. Table 1 shows the distribution 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 into a genome viewer.

Table 1
Counts of *E. coli* ORFs for each annotation category

Total ORFs				
135,097				
Coding (c_CDS)	Noncoding (nc_*)			
4460	130,637			
	Noncoding intergenic (nc_intergenic)	Noncoding overlapping with a genomic feature (nc_ovp_*)		
	18,318	112,319		
		On the same strand (nc_ovp_same-x)	On the opposite strand (nc_ovp_opp-x)	
		47,880	64,439	
			With x standing for:	
		45,053	CDS	62,354
		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 to characterize 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 h. 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 into 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 4 shows the two DNA strands of a genomic section of *E. coli* represented by the genome viewer IGV [27] 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 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

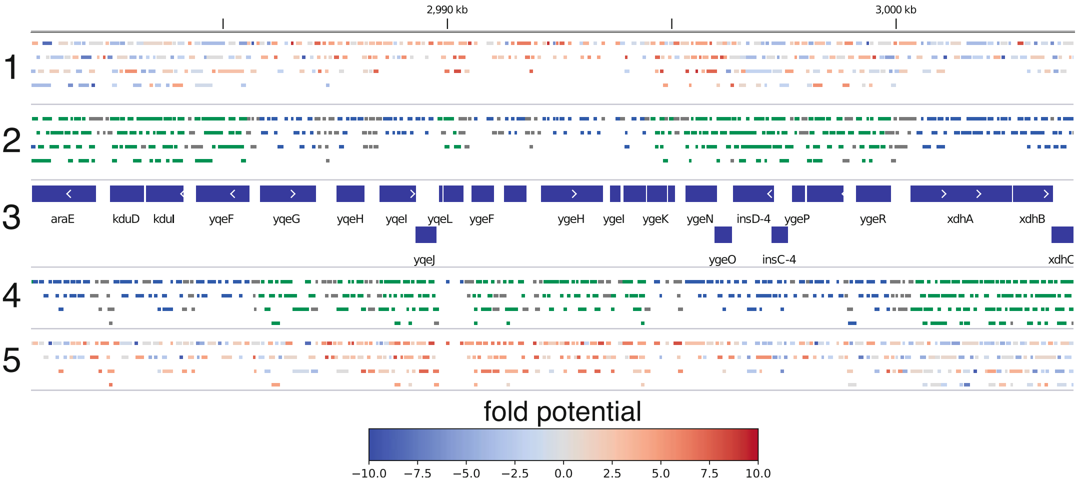

Fig. 4 Screenshot of a genomic section of *E. coli* represented by IGV. Genomic features present in the original GFF file (CDS in this example) are represented with blue boxes in the middle of the figure (panel 3). Panels 2 and 4 represent the noncoding ORFs identified by ORFtrack in the positive and negative strands, respectively. They are colored according to their annotation category (gray, 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 that 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

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 recalls the observation made by Bartonek et al. [30], 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 are a vast number 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.

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"
```

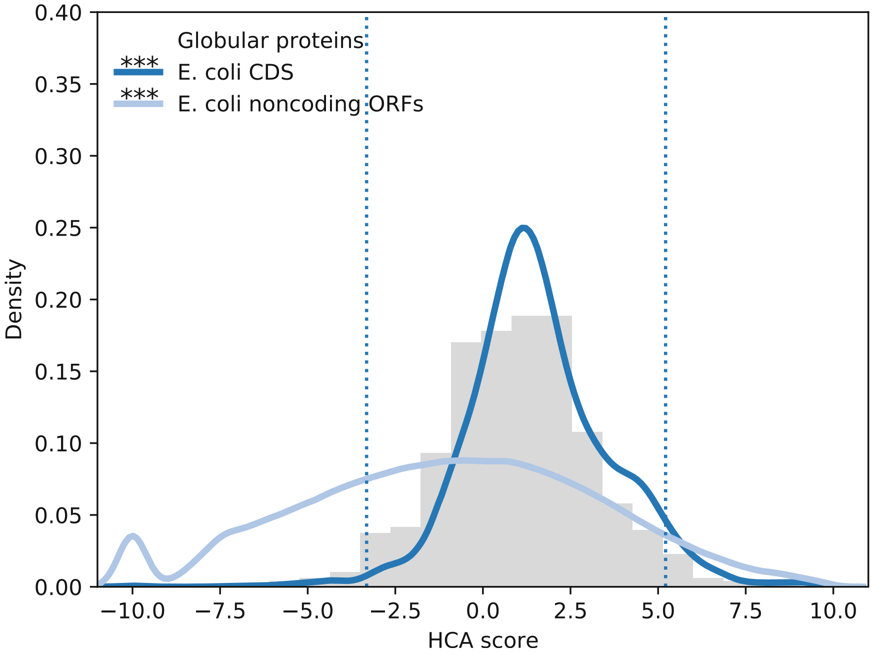

Fig. 5 Distribution of the HCA scores calculated for the CDS and the noncoding ORFs of *E. coli* (dark blue and light blue curves, respectively). The HCA score distribution of the set of globular proteins is represented by the gray histogram. Dotted black lines delineate the boundaries of the low, intermediate, and high HCA score bins so that 95% of the globular proteins fall into the intermediate HCA score bin. Each distribution is compared with that of the globular protein set with a Kolmogorov-Smirnov test. Asterisks on the plot denote level of significance: *** $<$ 0.001

Figure 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 [24]. The fold potential distribution of the CDS is clearly different from the one of the noncoding sequences (KS test, $P = 9.9 \times 10^{-18}$). The CDS is 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 most of them (~64%) exhibit similar HCA scores to globular proteins, revealing an important potential of foldable peptides, in line with the observation made in Fig. 4.

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 dramatic 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

1. M_musculus.fna.
2. M_musculus.gff (downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=mus+musculus).
3. E_coli.fna.
4. E_coli.gff (downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=e+coli).
5. H_volcanii.fna.
6. H_volcanii.gff. (downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=haloferax+volcanii).
7. D_melanogaster.fna.
8. D_melanogaster.gff (downloadable at https://www.ncbi.nlm.nih.gov/genome/?term=drosophila+melanogaster).

3.2.2 Annotation of ORFs of M. musculus *with ORFtrack*

In order to reduce the execution time (around 64 h on a single CPU), we recommend running ORFtrack on a cluster. The following command displays all the "seqid" values 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
```

Extracting all annotated ORFs with ORFget takes around 3 h on a single CPU and generates a 7.5 GB 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 20,000 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 20,000 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 *E. coli*. Consequently, in the next section, ORFold will be applied to a set of 20,000 randomly selected noncoding ORFs extracted from the complete set of mouse noncoding ORFs.

3.2.3 Extraction and Writing of the ORFs and CDS of M. musculus *with ORFget*

Definition of a Minimal Subset Size to Characterize the Fold Potential and Structural Properties of Noncoding ORFs

Extraction and Writing of the Amino Acid Sequences of a Dataset of 20,000 Noncoding ORFs

The following instruction allows the extraction of a subset of 20,000 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 *M. 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 20,000 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 20,000 noncoding ORFs and the 92,473 mouse isoforms (around 40 h 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 20,000 noncoding ORFs and all CDS of *H. volcanii*, *E. coli*, and *D. 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_melanogas-
ter_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
```

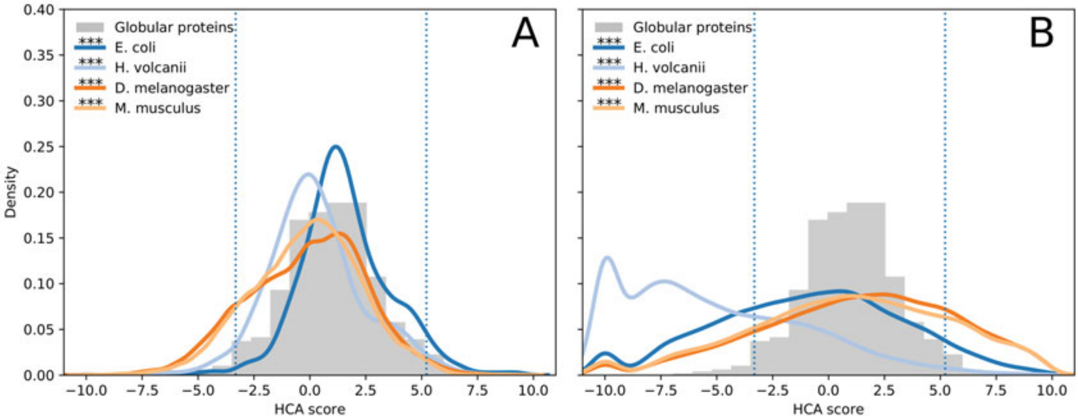

Fig. 6 (**a**) Distribution of the HCA scores calculated for the CDS of *E. coli*, *H. volcanii*, *D. melanogaster*, and *M. musculus* (dark blue, light blue, dark orange, and light orange curves, respectively). (**b**) Distribution of the HCA scores calculated for the noncoding ORFs of *E. coli*, *H. volcanii*, *D. melanogaster,* and *M. musculus* (dark blue, light blue, dark orange, and light orange curves, respectively). The HCA score distribution of the globular proteins is presented with the gray histogram. Each distribution is compared with the one of the globular proteins set with a Kolmogorov-Smirnov test. Asterisks on the plot denote the level of significance: *** < 0.001

```
D_melanogaster_noncoding.tab mouse_noncoding.tab -names "E.
coli" "H. volcanii" "D. melanogaster" "M. musculus"
```

Figure 6 shows, for the four species, the HCA score distributions of the corresponding CDS (Fig. 6a) and noncoding ORFs (Fig. 6b). Although the fold potential distributions of the CDS display slight variations among the four species, the vast majority (more than 85%) exhibit 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 *H. 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 *H. 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.

3.3 Probing the Fold Potential of a Set of Mouse Noncoding ORFs Shown to Be Pervasively Translated

Recently, Ruiz-Orera et al. [1] revealed with ribosome profiling experiments the translation of 721 ORFs in mouse lncRNAs (i.e., translated lncRNA-ORFs). They are not conserved across neighboring species nor subjected to selective pressure. The authors propose them as intermediates between noncoding ORFs and de novo genes [1]. This prompts us to ask whether their corresponding peptides display specific structural properties compared to peptides 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 20,000 randomly selected noncoding ORFs defined in Subheading 3.2. The amino acid sequences of all translated products identified in Ruiz_Orera et al. [1] (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, 20,000 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 20,000 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 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, $P = 5.9 \times 10^{-6}$, 4.8×10^{-6}, and

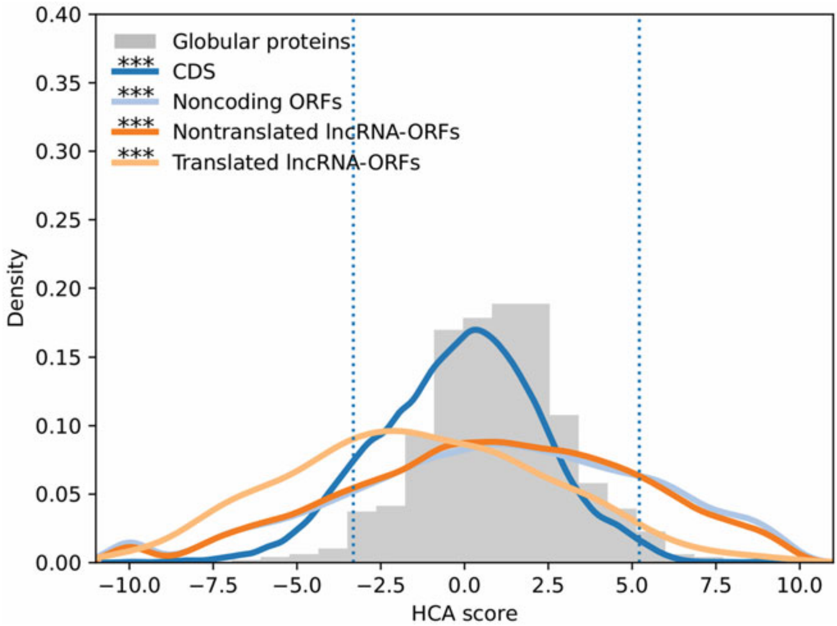

Fig. 7 Distribution of the HCA scores calculated for the CDS, the 20,000 noncoding ORFs, the 2000 nontranslated lncRNA-ORFs, and the 721 translated lncRNA-ORFs of *M. musculus* (dark blue, light blue, dark orange, and light orange curves, respectively). The HCA score distribution of the set of globular proteins is presented with the gray histogram. Each distribution is compared with that of the globular proteins with a Kolmogorov-Smirnov test. Asterisks on the plot denote the level of significance: *** < 0.001

2.4×10^{-6} 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. [31] 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.

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 a 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 [11, 32–37]. 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 the feature “CDS” usually matches 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.
2. Notice that the following instructions will lead to the same result:

```
 > orfget -fna E_coli.fna -gff mapping_orf_E_coli.gff -fea-
tures_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 20,000 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
```

References

1. 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
2. Chen J, Brunner A-D, Cogan JZ et al (2020) Pervasive functional translation of noncanonical human open reading frames. Science 367: 1140–1146
3. 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
4. Li J, Liu C (2019) Coding or noncoding, the converging concepts of RNAs. Front Genet 10: 496
5. Slavoff SA, Mitchell AJ, Schwaid AG et al (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nat Chem Biol 9:59
6. Prabakaran S, Hemberg M, Chauhan R et al (2014) Quantitative profiling of peptides from RNAs classified as noncoding. Nat Commun 5: 5429
7. Samayoa J, Yildiz FH, Karplus K (2011) Identification of prokaryotic small proteins using a comparative genomic approach. Bioinformatics 27:1765–1771
8. Hobbs EC, Fontaine F, Yin X, Storz G (2011) An expanding universe of small proteins. Curr Opin Microbiol 14:167–173
9. Eguen T, Straub D, Graeff M, Wenkel S (2015) MicroProteins: small size–big impact. Trends Plant Sci 20:477–482
10. Deng Y, Bamigbade AT, Hammad MA et al (2018) Identification of small ORF-encoded peptides in mouse serum. Biophys Rep 4: 39–49
11. Wang S, Mao C, Liu S (2019) Peptides encoded by noncoding genes: challenges and perspectives. Signal Transduct Target Ther 4: 1–12
12. Carvunis A-R, Rolland T, Wapinski I et al (2012) Proto-genes and de novo gene birth. Nature 487:370–374
13. 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
14. Tretyachenko V, Vymĕtal J, Bednárová L et al (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep 7:1–9
15. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715–718
16. Neme R, Amador C, Yildirim B et al (2017) Random sequences are an abundant source of bioactive RNAs or peptides. Nat Ecol Evol 1: 1–7
17. Faure G, Callebaut I (2013) Comprehensive repertoire of foldable regions within whole genomes. PLoS Comput Biol 9:e1003280
18. Faure G, Callebaut I (2013) Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. Bioinformatics 29: 1726–1733
19. Bitard-Feildel T, Callebaut I (2018) HCAtk and pyHCA: A toolkit and python API for the hydrophobic cluster analysis of protein sequences. bioRxiv 249995
20. 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
21. 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. J Mol Biol 342:345–353
22. 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

23. Rousseau F, Schymkowitz J, Serrano L (2006) Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol 16: 118–126
24. 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
25. Erdős G, Dosztányi Z (2020) Analyzing protein disorder with IUPred2A. Curr Protoc Bioinformatics 70:e99
26. Dosztányi Z (2018) Prediction of protein disorder based on IUPred. Protein Sci 27: 331–340
27. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
28. Bitard-Feildel T, Callebaut I (2017) Exploring the dark foldable proteome by considering hydrophobic amino acids topology. Sci Rep 7: 1–13
29. Mészáros B, Simon I, Dosztányi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5: e1000376
30. Bartonek L, Braun D, Zagrovic B (2020) Frameshifting preserves key physicochemical properties of proteins. Proc Natl Acad Sci U S A 117:5907–5912
31. 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
32. Yin X, Jing Y, Xu H (2019) Mining for missed sORF-encoded peptides. Expert Rev Proteomics 16:257–266
33. 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
34. Yadav M, Jhunjhunwala S, Phung QT et al (2014) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature 515:572–576
35. Sendoel A, Dunn JG, Rodriguez EH et al (2017) Translation from unconventional 5′ start sites drives tumour initiation. Nature 541:494–499
36. Barbosa C, Peixeiro I, Romão L (2013) Gene expression regulation by upstream open reading frames and human disease. PLoS Genet 9:e1003529
37. 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. Mol Genet Genomic Med 5:261–268