fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences

Paul M Harrison Corresp. 1

1 Dept. of Biology, McGill University, Montreal, QC, Canada

Corresponding Author: Paul M Harrison
Email address: paul.harrison@mcgill.ca

Compositionally-biased (CB) regions in biological sequences are enriched for a subset of sequence residue types. These can be shorter regions with a concentrated bias (*i.e.,* those termed 'low-complexity'), or longer regions that have a compositional skew. These regions comprise a prominent class of the uncharacterized 'dark matter' of the protein universe. Here, I report the latest version of the fLPS package for the annotation of CB regions, which includes added consideration of DNA sequences, to label the eight possible biased regions of DNA. In this version, the user is now able to restrict analysis to a specified subset of residue types, and also to filter for previously annotated domains to enable detection of discontinuous CB regions. A 'thorough' option has been added which enables the labelling of subtler biases, typically made from a skew for several residue types. In the output, protein CB regions are now labelled with bias classes reflecting the physico-chemical character of the biasing residues. The fLPS 2.0 package is available from: https://github.com/pmharrison/flps2 or in a supplementary file of this paper.

fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences

Paul M. Harrison

Department of Biology,

McGill University,

Montreal, QC,

Canada.

Corresponding author:

Paul Harrison

Email: paul.harrison@mcgill.ca

Abstract: Compositionally-biased (CB) regions in biological sequences are enriched for a subset of sequence residue types. These can be shorter regions with a concentrated bias (*i.e.,* those termed 'low-complexity'), or longer regions that have a compositional skew. These regions comprise a prominent class of the uncharacterized 'dark matter' of the protein universe. Here, I report the latest version of the fLPS package for the annotation of CB regions, which includes added consideration of DNA sequences, to label the eight possible biased regions of DNA. In this version, the user is now able to restrict analysis to a specified subset of residue types, and also to filter for previously annotated domains to enable detection of discontinuous CB regions. A 'thorough' option has been added which enables the labelling of subtler biases, typically made from a skew for several residue types. In the output, protein CB regions are now labelled with bias classes reflecting the physico-chemical character of the biasing residues. The fLPS 2.0 package is available from: https://github.com/pmharrison/flps2 or in a supplementary file of this paper.

Introduction

Biological sequences, despite being made from fixed alphabets of residues, demonstrate a wide diversity of sequence compositions. In particular, these sequences can be compositionally biased (CB) for a subset of the residue alphabet. For example, the protein sequence tract EDEEKDDELEIEEDEEDDDEDEDED is biased for E and D (glutamate and aspartate). If these tracts are sufficiently biased or repetitive over a short stretch, then they are termed 'low-complexity'. Also, one can have longer tracts that exhibit a milder compositional skew. In between, there are a continuum of CB cases

(Harrison 2006; Harrison & Gerstein 2003). In proteins, CB regions are linked to distinct biophysical states such as intrinsic disorder, and to cell-structural proteins, fibrous proteins, and functional amyloids and prions (Harbi & Harrison 2014; Harrison 2006), and to the formation of intracellular biomolecular condensates or membraneless organelles (Gomes & Shorter 2019). They also comprise part of the protein 'dark matter' that remains largely un- or under-characterized (Harrison 2018); indeed, some CB dark matter is not assignable as intrinsically disordered or structured, and may give us clues to as yet unknown biophysical protein states (Harrison 2018).

Several programs to annotate CB regions—and in particular, low-complexity (LC) regions—have been developed. These include SIMPLE (Hancock & Armstrong 1994), SEG (Wootton & Federhen 1996), CAST (Promponas et al. 2000), 0j.py (Wise 2001), ScanCom (Nandi et al. 2003), CARD (Shin & Kim 2005), BIAS (Kuznetsov & Hwang 2006), LCD-Composer (Cascarina et al. 2021) and LPS / fLPS (Harrison 2006; Harrison 2017; Harrison & Gerstein 2003). SEG annotates LC sequences by performing a scan using thresholds for sequence entropy and a fixed window length. It is used for masking LC sequences as part of the BLAST sequence alignment package (Altschul et al. 1997). Such masking is sometimes needed to avoid false inference of similarity by evolutionary descent (since these simpler sequences can arise independently multiple times during evolution quite easily). CAST annotates LC sequence by aligning to homopeptides of the twenty amino acids (Promponas et al. 2000). LCD-Composer uses a measure of amino-acid dispersion to characterise low complexity (Cascarina et al. 2021). The LCT web server analyzes the low-complexity and 'repeatability' of proteins sequences with a graphical output (Mier & Andrade-Navarro 2020). Two other servers LCReXXXplorer and

PLaToLoCo combine the results of multiple programs to graphically display LC regions (Jarnot et al. 2020; Kirmitzoglou & Promponas 2015). The LPS algorithm used binomial probability to check for low-probability sequence regions, and was further developed into the fast algorithm fLPS, which can annotate the TrEMBL database in <1 hour (Harbi et al. 2011; Harrison 2017; Harrison & Gerstein 2003). This algorithm has been applied successfully to the analysis of prions and prion-like proteins, and protein 'dark matter' (An et al. 2016; An & Harrison 2016; Harbi et al. 2011; Harrison et al. 2007; Harrison 2018; Harrison 2020; Su & Harrison 2019; Su & Harrison 2020).

The fLPS program is especially useful for analyzing CB regions since: (i) it analyzes the full range of CB types (from low complexity to milder compositional skews); (ii) it characterizes both single- and multiple-residue biases; (iii) it does not require the specification of residue types by the user (although this option has now been added); (iv) it considers the differing background frequencies of individual residue types; (v) it is faster than the commonly used SEG algorithm (Harrison 2017). fLPS has been applied to, for example, the identification of transactivation domains {Arnold, 2018 #33}, to analysis of the conservation of low-complexity regions in prokaryotes {Ntountoumi, 2019 #34}, analysis of low-complexity regions in stress granules {Zhu, 2020 #35}, and the delineation of domains in kinetochore proteins {Cortes-Silva, 2020 #36} and in the PRR19 protein that functions in meiotic crossing over {Bondarieva, 2020 #37}, as well as in studies of prion-like protein evolution {Harrison, 2020 #26}{An, 2016 #3;An, 2016 #2}{Su, 2019 #39;Su, 2020 #38}. Here, the latest fLPS 2.0 package is reported. In this package, the program flow has been modified to add consideration of DNA sequences; also, the user can specify subsets of residue types and existing domain annotations to filter from

sequences, in order to discover discontinuous biased regions. The baseline precision of the algorithm can be adjusted to discover more mildly biased regions that may have biological significance. Examples of fLPS 2.0 application are presented and discussed.

Methods

Implementation

The fLPS 2.0 package is written in standard C. The name 'fLPS' stands for fast LPS, where LPS stands for 'Low Probability Subsequences'. The package comprises the source code and executables compiled for MacOSX and Linux. There is the fLPS program itself, plus two accessory programs: *CompositionMaker*, which can be used to calculate background residue compositions; and *DomainFilter*, which is used to either excise or mask previously annotated domains (such as those with known protein structure, see section immediately below). Each of the programs works on input files of any size in standard FASTA format. The package is available at https://github.com/pmharrison/flps2 and in Supplementary File 1.

Algorithm and new added features

The program fLPS works through a process of binomial probability (P-value) minimization, as described in detail previously (Harrison 2017). There are four main steps that are summarized in Figure 1 at the top of the figure: *(i) QUICK SCAN; (ii) MINIMIZE; (iii) MERGE; (iv) OUTPUT.* At the end of the process, single-, and multiple-residue LPSs, are output if they are below the user-specified P-value threshold, or default threshold. Biased regions are labelled with a *bias signature* which is a list of the biasing residues in

order of bias precedence delimited with curly brackets. At each of these stages, efficiency measures are taken to avoid or delay probability calculations unless/until they are necessary (Harrison 2017).

The following are the main new options added to the fLPS code:

(i) *Precision of the calculation:* In the initial *QUICK SCAN* step, by default (the '*–z fast*' option), windows with a P-value below the baseline threshold of 0.001 are considered. Also, the windowing along the sequence proceeds with a step size = 3 residues (Figure 1). This means that some regions that are made from biases for a larger number of residue types might be missed; also, short regions with a milder bias that might have biological significance could sometimes be overlooked. Therefore, options for the base-line precision of the program have been added. If '*–z medium'* is specified, the base-line P-value threshold is set to 0.01, with a windowing step size = 2. For the most precise option '*–z thorough'*, the base-line P-value is 0.1 and the step size = 1 (Figure 1). However, these latter two options can produce a huge amount of output for larger databases, so they should be applied to such databases with caution.

(ii) *DNA analysis:* DNA sequences can be specified using the *–n* option. By default, each of the four bases A, G, C and T has equal background probability.

(iii) *Domain filtering:* Using the *DomainFilter* accessory program, previously annotated domains can be filtered in either of two ways*, i.e.,* either 'excised' or 'masked'. When 'excised' is specified, *DomainFilter* outputs shorter sequences, with the domain sequences removed. The 'masked' option outputs sequences with the domains masked with Xs. The positions of the excised or masked

domains are labelled on the name line of the sequences in the FASTA-format file. When the FASTA-format output file from *DomainFilter* is used as input for fLPS bias annotations, the domain positions appear in the fLPS output if the -option *–D* is specified.

(iv) *Restriction lists:* With the *–r* option, the user can specify a subset of residue types, *e.g.*, only negatively-charged amino acids (E, D), or the six-membered aromatics (F, Y, W).

Further option additions include: An option (*–O)* option to specify a prefix for a unique output filename that also contains the parameters used in running the fLPS program; a *'–o oneline'* output option, wherein the results for each sequence are listed in a single-line summary; a *–k* option to ignore the unknown residues in calculations ('X' for proteins and 'N' for DNA). The output has also been updated to include further new features. A calculation of the enrichment of the biasing residues in the output LPSs has been added, which is the proportion of biasing residues in the LPS divided by the total expected background frequencies of the biasing residues. To enable quicker characterization of bias trends in a data set, '*bias class*' labels are now featured for both protein and DNA sequences. For proteins, these labels are derived from the Taylor amino-acid classification Venn diagram, with some additional categories (Taylor 1986). The applicable class label that has the smallest membership is picked, when assigning these. For DNA, these labels represent the eight possible compositional biases: *{A}|{T}, {G}|{C}, {AT}, {GC}, {AC}|{GT}, {AG}|{CT}, {ATC}|{ATG}* and *{ACG}|{CGT}* (discussed below).

For better annotation of short low-complexity regions, trimming of LPSs of minimum window length is now employed. That is, if possible, residues are sheared off both ends

of the minimum-length LPS if they do not contribute to the bias. This improves the annotation of ~6-8% of LPSs, in trials on the *S. cerevisiae* S288C proteome (downloaded from UniProt reference proteomes (Boeckmann et al. 2003)) using a variety of parameters.

Example data

The UniProt canonical reference human and budding yeast (*S. cerevisiae* strain 288C) proteomes were downloaded from www.uniprot.org in January 2021 (Boeckmann et al. 2003). The human proteome was cross-referenced with the InterPro list of domain annotations downloaded from http://ebi.ac.uk/interpro, to make a list of human proteins that contain the RRM RNA-binding domain (Blum et al. 2021).

Human promoter data was obtained from the EPD eukaryotic promoter database (Schmid et al. 2004). These were a set of representative promoters (one per gene) defined by the EPD. Sequences spanning from –999 to +100 around the transcription start site were analysed.

Prion-like regions

Prion-like regions were annotated for the human proteome using the PLAAC program with default parameters (Lancaster et al. 2014).

Results

Using the new domain-filtering and restriction list options of fLPS: application to analysis of human RNA-binding proteins

178 It is often advantageous to restrict CB annotation to a subset of residues to enable
179 easy counting of different types of bias region. Users are now able to restrict their bias
180 annotation using a 'restriction list' specified with the *–r* option of fLPS (Figure 2a). Also, it
181 is possible that certain proteins have discontinuous CB regions, *i.e.,* the CB regions may
182 have small, structured domains embedded in them, or they may be comprised of the loop
183 regions within a single protein domain. To enable discovery of such discontinuous CB
184 regions, the *DomainFilter* program can be used to excise or mask domains or domain
185 parts before using the fLPS program (Figure 2a). These two options were combined in
186 analyzing the CB regions of human RNA-binding proteins, specifically those containing
187 the RRM RNA-binding domain (Figure 2). The RRM domain is used in eukaryotes to bind
188 RNA during diverse cellular processes, and is typically associated with intrinsically-
189 disordered regions (Su & Harrison 2020). After applying *DomainFilter* to excise Pfam
190 protein domain annotations (Mistry et al. 2021), the main single-residue biases were
191 assessed with an initial run of fLPS (those having >50 cases); thereafter, a final run of
192 fLPS used a restriction list based on these main single-residue biases to enable better
193 counting of bias types.

194 There are 206 RRM-domain-containing human proteins in this protein data set.
195 The most common multiple- and single-residue biases involve arginine, serine, proline
196 and glycine, and are associated mostly with 'mixed', 'polar', 'small' and 'charged' bias
197 classes; glutamine and asparagine biases, which are associated typically with prion-like
198 domains, are only of middling abundance (Figure 2b-d). Indeed, although prion-like
199 domains are often cited as being associated with RNA-binding proteins, in this case they
200 only occur in ~1 in 6 RRM-containing proteins, as judged by the PLAAC program

(Lancaster et al. 2014) (Suppl. Figure 1, 32/206 (15.5%) have PLAAC LLR scores ≥15.0, and 35/206 (17.0%) have PLAAC PRD scores >15.0). These PLAAC prion-like regions arise despite only moderate asparagine and glutamine frequencies and are thus substantially dependent on other residues that are common in prion-forming domains, such as tyrosine, glycine and serine, which are common biases in the RRM-containing proteins (Figure 2b-d). In Figure 2e, the human BOLL 'protein boule-like' is presented as an example of a discontinuous CB domain around an RRM domain.

Increased precision with the –z option

As described in *Methods*, the *–z* option can be used to increase the precision of the initial scanning by the fLPS algorithm for compositional deviations (Figure 1). Two examples of the effects of this option are illustrated (Figure 3). Multi-protein bridging factor MBF1 is a transcriptional coactivator that promotes GCN4-dependent transcriptional activity by bridging between the DNA-binding areas of GCN4 and TATA-binding protein. The default fLPS settings detect a mild bias for positively-charged residues *{KR}*, which becomes a stronger bias comprised of further biasing residues *{KRQGANVSP}* when the 'thorough' option is applied. Also, a region weakly biased for polar residues *{TDN}* appears (Figure 3a). These biases are likely linked to DNA and protein interactions within complexes. The second example is the Shadoo protein from human (Figure 3b). Shadoo is a member of the prion-protein (PrP) family that has demonstrated some neuroprotective behaviour (Westaway et al. 2011). Like PrP, the protein that underlies prion diseases, it contains CB and intrinsically-disordered regions. Here, the major CB annotations are stable when *–z thorough* is specified, but additional mildly biased regions are detected,

one of which corresponds to a signal peptide, the other an area bridging between and intrinsically disordered region and a pro-peptide (Figure 3b).

These examples demonstrate three effects of increasing the precision of the initial compositional scanning: *(i)* mildly biased tracts are detected that can be quite short and that may have biological significance; *(ii)* further bias detail is sometimes added to CB regions, decreasing the binomial P-value; *(iii)* tracts with a bias made from several residues and that were previously not detectable (such as the {AWLC} tract in Figure 1(b)) become evident. In aggregate, these three effects increase the ability of the program to delineate compositionally-defined domains in proteins. As shown in Table 1 detailing analysis of the *S. cerevisiae* proteome, a significant number of further multiple-residue CB regions are detected, even for smaller P-value thresholds (such as P≤1e-09). Since one of the tracts in Figure 1(b) corresponds to a signal peptide, the correspondence between signal peptide positions in the *S. cerevisiae* proteome and CB regions was also examined (Table 2). The number of signal peptides corresponding to CB regions increases to a highest value of ~60% with a P≤1e-03 bias P-value threshold. The results are generally in line with a previous analysis of sequence complexity in signal peptides, where 24% of residues of signal peptides in analyzed data sets were labelled part of low-complexity tracts by SEG {Wong, 2010 #42}{Wootton, 1996 #5}.

Analysis of DNA sequence

DNA sequences can be analyzed by specifying the *–n* option; by default, each base is expected with equal background probability. In total, there are 40 different possible biases (Figure 4a). These can be segregated into eight bias classes for DNA

(Figure 4a). The last two of these bias classes, *{ATC}|{ATG}* and *{ACG}|{CGT}*, correspond to strand-specific depletions of single bases, *i.e., {ATC}|{ATG}* indicates a strand-specific lack of C or G. An example of a DNA CB region from a human promoter is illustrated (Figure 4b). To illustrate its application to DNA, we used the fLPS program to examine bias trends in a representative data set of human promoters taken from the EPD database (Figure 4c). Interestingly, the tri-base *{ATC}|{ATG}* and *{ACG}|{CGT}* bias classes are almost as prevalent as the two-base *{AC}|{GT}* bias class.

Discussion

The known protein universe contains much 'dark matter', some structured, some intrinsically disordered, some not assignable as either (Harrison 2018). The present fLPS package helps to address one aspect of this protein dark matter, which is that it often has unusual amino-acid composition, the structural properties of which have yet to be characterized. As examined above for human RRM-domain-containing proteins, some of this dark matter may be discontinuous CB regions that have smaller structured domains embedded in them. Such discontinuity might also be possible within a protein domain, *e.g.,* in the loops of a transmembrane domain. fLPS can also be used to assess whether such compositional biases are unusual relative to the background proteome composition of a particular organism or clade, or are part of organism- / clade-specific trend. CB domains, such as the proline-rich region in PRR19 protein {Bondarieva, 2020 #37} that functions in meiotic crossing-over, may have specific functional relevance. Parsing proteins into subdomains, including milder CB domains found with an increased baseline thoroughness (the –z option), may help in the generation of experimental constructs and

further hypotheses for experiments. They are also useful for studying proteome-wide trends to gain more general functional or evolutionary insights (*e.g.*, refs. {Ntountoumi, 2019 #34}{Su, 2020 #38}). Varying the parameters (in particular –*m*, –*M*, –*t*, and the new parameters –*r* and –*z*) can help to delineate possible biologically meaningful subdomains in larger biased tracts or within intrinsically disordered regions, such as in prion-forming proteins {Harrison, 2017 #22}.

The fLPS package program flow was modified to accommodate the option of analyzing DNA sequences. I applied this option to a set of representative human promoters, as an example. Beyond the standard conception of DNA bias as either {GC} or {AT}, substantial tracts of other possible biases were observed, including strand-specific dearths of single bases (*i.e.,* the bias classes *{ATC}|{ATG}* and *{ACG}|{CGT}*). It would be interesting to investigate experimentally whether such DNA CB domains have a general biological significance. To my knowledge, there is not a currently available program that delineates all of the possible biased domains of DNA in this way (other programs, such as Dustmasker {Morgulis, 2006 #40} or TANTAN {Frith, 2011 #41}, are designed to tackle the problem of avoiding spurious alignments, which is not what fLPS is designed for.)

Intrinsically disordered regions (IDRs) in proteins were initially discovered as long stretches of amino acids in proteins that remain unfolded under physiological conditions [1, 2]{Uversky, 2002 #28}. Compositional bias or ‘low complexity’ is a characteristic feature of intrinsically disordered regions (IDRs), although there is substantial overlap in sequence complexity values between IDRs and ordered regions {Pedro Romero, 2001 #27}. Also, different definitions of sequence complexity or compositional bias have

different degrees of linkage to disorder or order, with tandem-repeat tracts more likely to encode ordered regions {Mier, 2020 #29}. Because of this link, fLPS 2.0 may be useful for the characterization of subdomains in intrinsically-disordered proteins. The boundaries of compositionally-defined domains may differ to those of IDRs, IDRs may be split into multiple compositionally-defined regions, or new algorithmic scenarios using the definition of compositionally defined domains may enable the annotation of further intrinsic disorder {Necci, 2021 #30}{Sirota, 2010 #31}{Tang, 2021 #32}.

Examples of running the package

A diverse choice of parameters is possible in running the fLPS 2.0 program. Here are some examples:

(1) *Annotating low-complexity regions:* For the specific task of annotating short CB regions of the sort termed 'low-complexity', the following parameters are suitable (with the yeast proteome 'yeast.fasta' as an example input file):

./fLPS –t1e-5 (or –t1e-6) –m5 –M25 –o long yeast.fasta

(2) *Analyzing for discontinuous CB domains:* Firstly, structured domains are excised from the sequences, then fLPS is run using the –D option:

./DomainFilter –D excised yeast.fasta > yeast.Dexcised.fasta

./fLPS –D yeast.Dexcised.fasta

(3) *Restricting biases to specific sets:* To analyze biases for just six-membered aromatic side-chain amino acids only (F, Y and W), using the yeast proteome background composition:

./CompositionMaker yeast.fasta (makes file 'yeast.fasta.COMPOSITION')

316 *./fLPS –dv –ooneline –c yeast.fasta.COMPOSITION –r FYW yeast.fasta*

317 Also specified are headers and footers (–*d*), oneline output format (–*o oneline*) and

318 verbose behaviour during runtime (–*v*).

319 *(4) Annotating longer biased regions with the thorough option:* To find longer biased

320 regions that have compositional skew the following options may be suitable:

321 *./fLPS –z thorough –t0.001 –M 1000 yeast.fasta*

322 *(5) DNA:* For DNA, the –n option is specified:

323 *./fLPS –dn DNA.example.fasta*

324 Here, headers and footers are also output (–*d* option).

325

326 *Conclusions*

327 The fLPS 2.0 package is a versatile package for annotating compositional biases,

328 either 'low-complexity' regions, or regions with milder or long-range compositional skew.

329 Users can now apply the package to DNA to identify all the possible DNA CB domains.

330 In addition to the unique features of fLPS listed at the end of the Introduction, utility is

331 gained from the added domain filtering, restriction list and precision options, which can

332 be combined to identify CB domains in support of experimental hypotheses. The package

333 is available from: https://github.com/pmharrison/flps2 or Supplementary File 1.

334

335 ## Figure Legends

336

337 **Figure 1: A schematic detailing the –*z* option to adjust the base-line precision of**

338 **the fLPS calculation.** At the top of the figure is a pipeline summarizing the basic fLPS

339 algorithm. Below that is detailed the effect of the –*z* option for adjusting the base-line

precision of the algorithm. In the *QUICK SCAN* stage, when the *–z thorough* option is specified, more windows are stored in accord with the higher base-line P-value (these are coloured green). Thus, there are more and longer search contigs (surrounded with yellow box) at the end of this stage.

Figure 2: Analysis of RRM RNA-binding domain proteins in the human proteome. **(a)** Human Pfam domain annotations (coloured boxes) are excised with *DomainFilter* and the biases are annotated using fLPS 2.0 with human proteome background composition. The most prevalent single-residue biases (occurring >50 times) were picked out (listed in part (c)) and used as a restriction list with the *–r* option. **(b)** A bar chart of the most prevalent multiple-residue bias signatures (that occur for any threshold ≥3 times). The data for four P-value thresholds are shown. **(c)** As in (b) except that single-residue biases are counted. **(d)** As in (b), except the bias classes are counted. The following bias classes do not occur: glx, tiny_polar, polar_aromatic, aliphatic, aromatic. **(e)** An example of a discontinuous biased region from human BOLL protein. The RRM domain (Pfam PF00076, **underlined bold**) is excised. A *{P}* CB region with P-value = 7.2e-9 is shown in *italics* with the P residues in red. There are also a *{Y}* CB region (P = 8.6e-6, residues in green) and a *{Q}* CB region (P = 2.8e-5, residues in blue). These go together to make a *{PYQ}* region of the same extent as the *{P}* region with P = 4.4e-13. Other Q and Y residues in this multiple-residue CB region are in bold.

Figure 3: Two examples of the effect of the *–z* option: (a) Multiprotein-bridging factor MBF1 from *S. cerevisiae*; **(b)** human Shadoo protein.

363 Single-residue CB regions are depicted as blue boxes and multiple-residue as green.
364 They are labelled with their biases and binomial P-values, and their endpoints. Intrinsic
365 disorder and other domain annotations are labelled in orange and grey respectively (and
366 are taken from UniProt (Boeckmann et al. 2003)). At the top of each panel are depicted
367 the annotation from the default *–z fast* option, below that the annotations after using the
368 *–z thorough* option, then at the bottom of each panel are the UniProt sequence
369 annotations.
370
371 **Figure 4: Analysis of DNA: (a)** Eight classes of bias are possible in DNA.
372 Complementary biases are arrayed above and below the line, *i.e.,* a bias on one strand
373 for *{GT}* (guanine and thymidine) corresponds to a bias for *{CA}* (cytidine and adenine)
374 on the complementary strand. Biases with the same colour are summarized with one of
375 the eight basic bias class labels (in the box at the bottom of the panel).
376 **(b)** An example of a biased region in human promoter DNA, for colipase CLPS_1.
377 The position in the promoter (downloaded from the EPD database (Schmid et al. 2004))
378 is indicated, along with the bias signature *{AC}*.
379 **(c)** Prevalences of the eight bias classes in human promoters downloaded from
380 the EPD (Schmid et al. 2004). Data for each of three bias P-value thresholds are shown
381 (P≤1.0e-06, P≤1.0e-09 and P≤1.0e-12). The total number of residues in CB regions for
382 each of these thresholds is summed and depicted in natural logarithmic scale. The
383 numeric values are labelled on the top of each bar.
384

385 ## References

386 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997.
387 Gapped BLAST and PSI-BLAST: a new generation of protein database search
388 programs. *Nucleic Acids Res* 25:3389-3402.
389 An L, Fitzpatrick D, and Harrison PM. 2016. Emergence and evolution of yeast prion and
390 prion-like proteins. *BMC Evol Biol* 16:24. 10.1186/s12862-016-0594-3
391 An L, and Harrison PM. 2016. The evolutionary scope and neurological disease linkage
392 of yeast-prion-like proteins in humans. *Biol Direct* 11:32. 10.1186/s13062-016-
393 0134-5
394 Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G,
395 Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork
396 P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA,
397 Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N,
398 Thomas PD, Tosatto SCE, Wu CH, Bateman A, and Finn RD. 2021. The InterPro
399 protein families and domains database: 20 years on. *Nucleic Acids Res* 49:D344-
400 D354. 10.1093/nar/gkaa977
401 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ,
402 Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M. 2003. The SWISS-
403 PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids*
404 *Res* 31:365-370.
405 Cascarina SM, King DC, Osborne Nishimura E, and Ross ED. 2021. LCD-Composer: an
406 intuitive, composition-centric method enabling the identification and detailed
407 functional mapping of low-complexity domains. *NAR Genom Bioinform* 3:lqab048.
408 10.1093/nargab/lqab048

409 Gomes E, and Shorter J. 2019. The molecular language of membraneless organelles. *J*
410 *Biol Chem* 294:7115-7127. 10.1074/jbc.TM118.001192
411 Hancock JM, and Armstrong JS. 1994. SIMPLE34: an improved and enhanced
412 implementation for VAX and Sun computers of the SIMPLE algorithm for analysis
413 of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* 10:67-
414 70. 10.1093/bioinformatics/10.1.67
415 Harbi D, and Harrison PM. 2014. Classifying prion and prion-like phenomena. *Prion* 8:pii:
416 27960.
417 Harbi D, Kumar M, and Harrison PM. 2011. LPS-annotate: complete annotation of
418 compositionally biased regions in the protein knowledgebase. *Database (Oxford)*
419 2011:baq031. 10.1093/database/baq031
420 Harrison LB, Yu Z, Stajich JE, Dietrich FS, and Harrison PM. 2007. Evolution of budding
421 yeast prion-determinant sequences across diverse fungi. *J Mol Biol* 368:273-282.
422 10.1016/j.jmb.2007.01.070
423 Harrison PM. 2006. Exhaustive assignment of compositional bias reveals universally
424 prevalent biased regions: analysis of functional associations in human and
425 Drosophila. *BMC Bioinformatics* 7:441. 10.1186/1471-2105-7-441
426 Harrison PM. 2017. fLPS: Fast discovery of compositional biases for the protein universe.
427 *BMC Bioinformatics* 18:476. 10.1186/s12859-017-1906-3
428 Harrison PM. 2018. Compositionally Biased Dark Matter in the Protein Universe.
429 *Proteomics* 18:e1800069. 10.1002/pmic.201800069
430 Harrison PM. 2020. Variable absorption of mutational trends by prion-forming domains
431 during Saccharomycetes evolution. *PeerJ* 8:e9669. 10.7717/peerj.9669

Harrison PM, and Gerstein M. 2003. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biol* 4:R40. 10.1186/gb-2003-4-6-r40

Jarnot P, Ziemska-Legiecka J, Dobson L, Merski M, Mier P, Andrade-Navarro MA, Hancock JM, Dosztanyi Z, Paladin L, Necci M, Piovesan D, Tosatto SCE, Promponas VJ, Grynberg M, and Gruca A. 2020. PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic Acids Res* 48:W77-W84. 10.1093/nar/gkaa339

Kirmitzoglou I, and Promponas VJ. 2015. LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences. *Bioinformatics* 31:2208-2210. 10.1093/bioinformatics/btv115

Kuznetsov IB, and Hwang S. 2006. A novel sensitive method for the detection of user-defined compositional bias in biological sequences. *Bioinformatics* 22:1055-1063. 10.1093/bioinformatics/btl049

Lancaster AK, Nutter-Upham A, Lindquist S, and King OD. 2014. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* 30:2501-2502. 10.1093/bioinformatics/btu310

Mier P, and Andrade-Navarro MA. 2020. Assessing the low complexity of protein sequences via the low complexity triangle. *PLOS ONE* 15:e0239154. 10.1371/journal.pone.0239154

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, and Bateman A. 2021. Pfam: The

protein families database in 2021. *Nucleic Acids Res* 49:D412-D419. 10.1093/nar/gkaa913

Nandi T, Dash D, Ghai R, C BR, Kannan K, Brahmachari SK, Ramakrishnan C, and Ramachandran S. 2003. A novel complexity measure for comparative analysis of protein sequences from complete genomes. *J Biomol Struct Dyn* 20:657-668. 10.1080/07391102.2003.10506882

Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, and Ouzounis CA. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* 16:915-922.

Schmid CD, Praz V, Delorenzi M, Perier R, and Bucher P. 2004. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res* 32:D82-85. 10.1093/nar/gkh122

Shin SW, and Kim SM. 2005. A new algorithm for detecting low-complexity regions in protein sequences. *Bioinformatics* 21:160-170. 10.1093/bioinformatics/bth497

Su TY, and Harrison PM. 2019. Conservation of Prion-Like Composition and Sequence in Prion-Formers and Prion-Like Proteins of Saccharomyces cerevisiae. *Front Mol Biosci* 6:54. 10.3389/fmolb.2019.00054

Su WC, and Harrison PM. 2020. Deep conservation of prion-like composition in the eukaryotic prion-former Pub1/Tia1 family and its relatives. *PeerJ* 8:e9023. 10.7717/peerj.9023

Taylor WR. 1986. The classification of amino acid conservation. *J Theor Biol* 119:205-218. 10.1016/s0022-5193(86)80075-3

Westaway D, Daude N, Wohlgemuth S, and Harrison P. 2011. The PrP-like proteins Shadoo and Doppel. *Top Curr Chem* 305:225-256. 10.1007/128_2011_190

Wise MJ. 2001. 0j.py: a software tool for low complexity proteins and protein domains. *Bioinformatics* 17 Suppl 1:S288-295. 10.1093/bioinformatics/17.suppl_1.s288

Wootton JC, and Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 266:554-571.

484

Figure 1

A schematic detailing the *–z* option to adjust the base-line precision of the fLPS calculation.

At the top of the figure is a pipeline summarizing the basic fLPS algorithm. Below that is detailed the effect of the *–z* option for adjusting the base-line precision of the algorithm. In the *QUICK SCAN* stage, when the *–z thorough* option is specified, more windows are stored in accord with the higher base-line P-value (these are coloured green). Thus, there are more and longer search contigs (surrounded with yellow box) at the end of this stage.

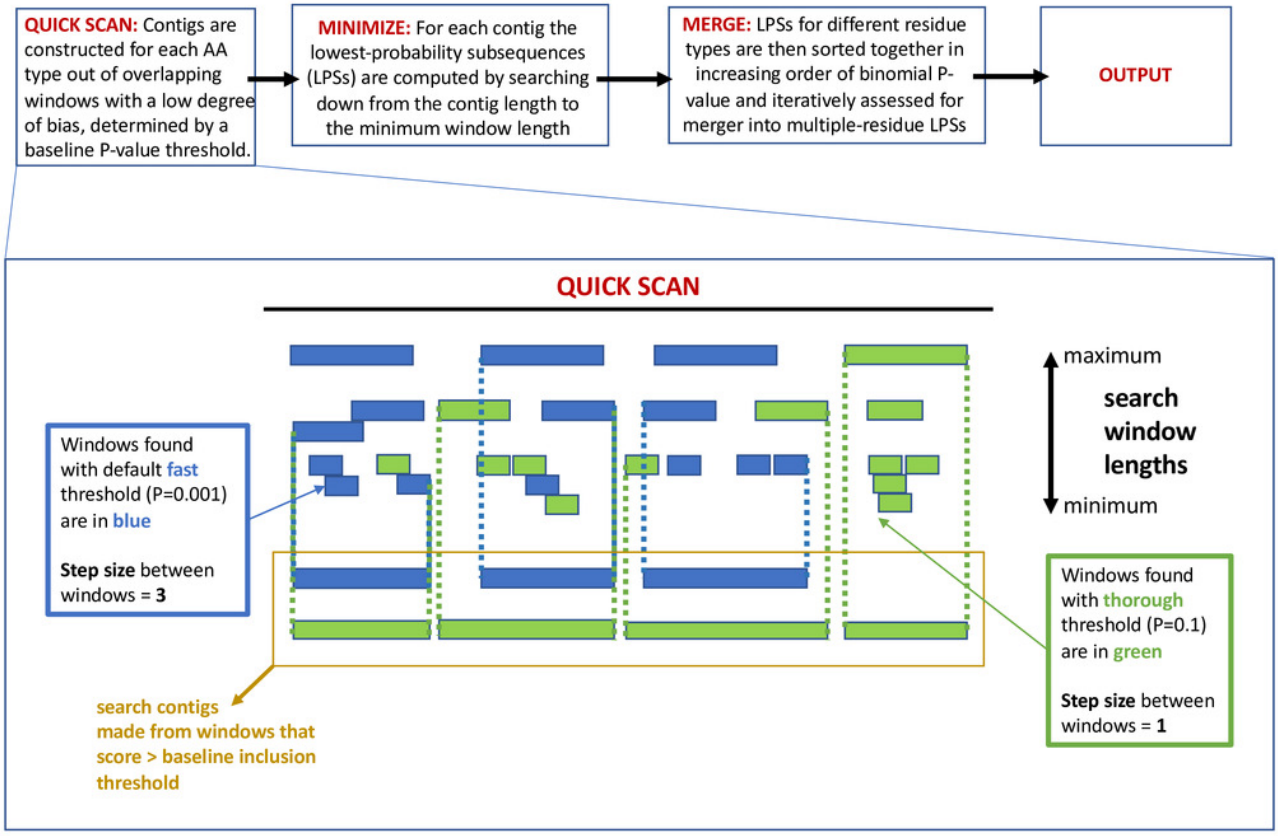

Figure 2

Analysis of RRM RNA-binding domain proteins in the human proteome.

(a) Human Pfam domain annotations (coloured boxes) are excised with *DomainFilter* and the biases are annotated using fLPS 2.0 with human proteome background composition. The most prevalent single-residue biases (occurring >50 times) were picked out (listed in part (c)) and used as a restriction list with the *–r* option. **(b)** A bar chart of the most prevalent multiple-residue bias signatures (that occur for any threshold ≥3 times). The data for four P-value thresholds are shown. **(c)** As in (b) except that single-residue biases are counted. **(d)** As in (b), except the bias classes are counted. The following bias classes do not occur: glx, tiny_polar, polar_aromatic, aliphatic, aromatic. **(e)** An example of a discontinuous biased region from human BOLL protein. The RRM domain (Pfam PF00076, **underlined bold**) is excised. A *{P}* CB region with P-value = 7.2e-9 is shown in *italics* with the P residues in red. There are also a *{Y}* CB region (P = 8.6e-6, residues in green) and a *{Q}* CB region (P = 2.8e-5, residues in blue). These go together to make a *{PYQ}* region of the same extent as the *{P}* region with P = 4.4e-13. Other Q and Y residues in this multiple-residue CB region are in bold.

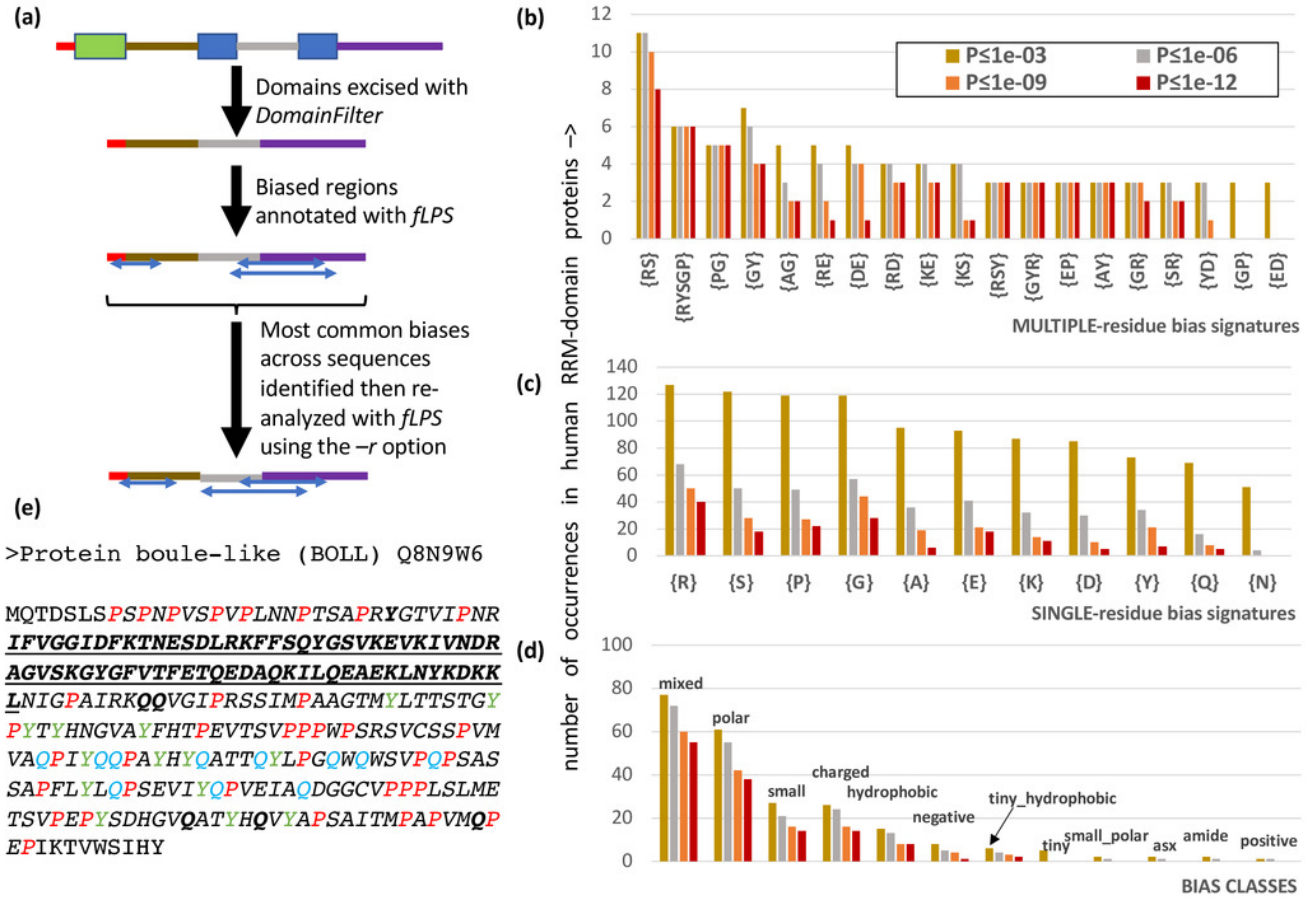
(a)
Domains excised with DomainFilter
Biased regions annotated with fLPS
Most common biases across sequences identified then re-analyzed with fLPS using the –r option
(b)
P≤1e-03
P≤1e-06
P≤1e-09
P≤1e-12
number of occurrences in human RRM-domain proteins -->
{RS}
{RYSGP}
{PG}
{GY}
{AG}
{RE}
{DE}
{RD}
{KE}
{KS}
{RSY}
{GYR}
{EP}
{AY}
{GR}
{SR}
{YD}
{GP}
{ED}
MULTIPLE-residue bias signatures
(c)
{R}
{S}
{P}
{G}
{A}
{E}
{K}
{D}
{Y}
{Q}
{N}
SINGLE-residue bias signatures
(d)
mixed
polar
small
charged
hydrophobic
negative
tiny_hydrophobic
tiny
small_polar
asx
amide
positive
BIAS CLASSES
(e)
>Protein boule-like (BOLL) Q8N9W6
MQTDSLSPSPNPVSPVPLNNPTSAPRYGTVIPNR
IFVGGIDFKTNESDLRKFFSQYGSVKEVKIVNDR
AGVSKGYGFVTFETQEDAQKILQEAEKLNYKDKK
LNIGPAIRKQQVGIPRSSIMPAAGTMYLTTSTGY
PYTYHNGVAYFHTPEVTSVPPPWPSRSVCSSPVM
VAQPIYQQPAYHYQATTQYLPGQWQWSVPQPSAS
SAPFLYLQPSEVIYQPVEIAQDGGCVPPPLSLME
TSVPEPYSDHGVQATYHQVYAPSAITMPAPVMQP
EPIKTVWSIHY

Figure 3

Two examples of the effect of the *–z* option: (a) Multiprotein-bridging factor MBF1 from *S. cerevisiae*; (b) human Shadoo protein.

Single-residue CB regions are depicted as blue boxes and multiple-residue as green. They are labelled with their biases and binomial P-values, and their endpoints. Intrinsic disorder and other domain annotations are labelled in orange and grey respectively (and are taken from UniProt (Boeckmann et al. 2003)). At the top of each panel are depicted the annotation from the default *–z fast* option, below that the annotations after using the *–z thorough* option, then at the bottom of each panel are the UniProt sequence annotations.

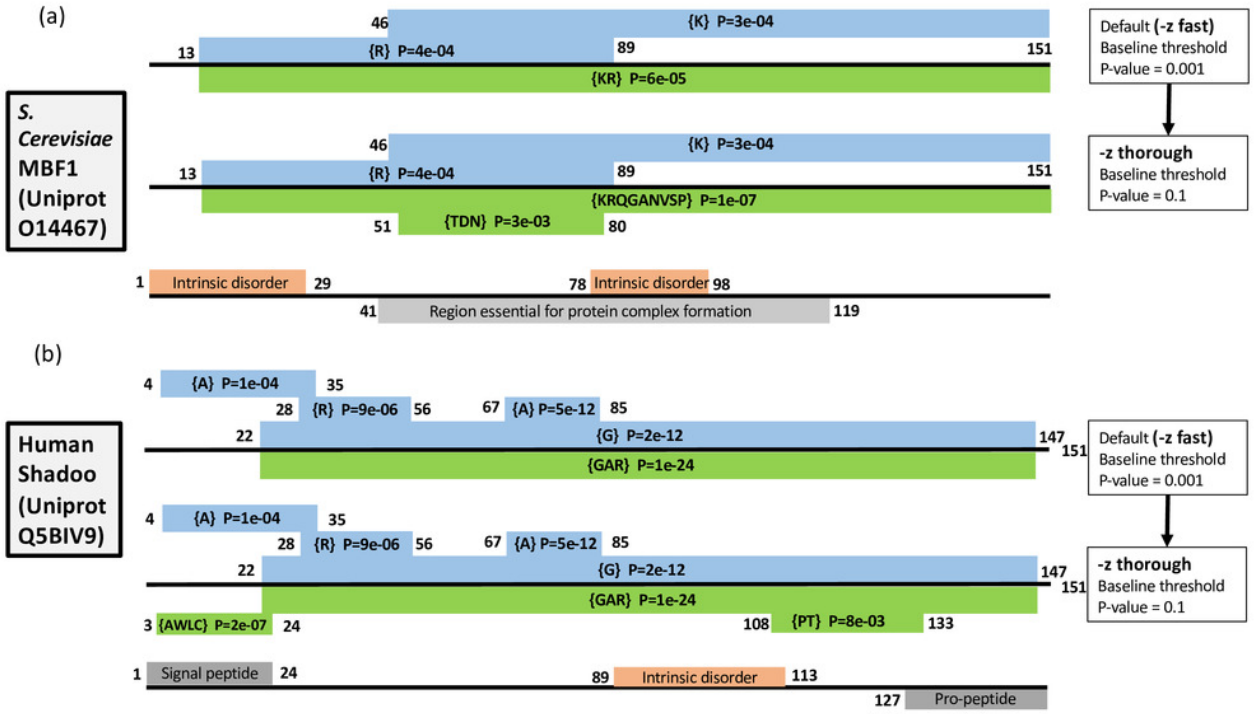
(a)
S. Cerevisiae MBF1 (Uniprot O14467)
46 {K} P=3e-04
13 {R} P=4e-04 89 151
{KR} P=6e-05
Default (-z fast) Baseline threshold P-value = 0.001
-z thorough Baseline threshold P-value = 0.1
{KRQGANVSP} P=1e-07
51 {TDN} P=3e-03 80
1 Intrinsic disorder 29 78 Intrinsic disorder 98
41 Region essential for protein complex formation 119
(b)
Human Shadoo (Uniprot Q5BIV9)
4 {A} P=1e-04 35
28 {R} P=9e-06 56 67 {A} P=5e-12 85
22 {G} P=2e-12 147 151
{GAR} P=1e-24
3 {AWLC} P=2e-07 24 108 {PT} P=8e-03 133
1 Signal peptide 24 89 Intrinsic disorder 113
127 Pro-peptide

Figure 4

Analysis of DNA

(a) Eight classes of bias are possible in DNA. Complementary biases are arrayed above and below the line, *i.e.,* a bias on one strand for *{GT}* (guanine and thymidine) corresponds to a bias for *{CA}* (cytidine and adenine) on the complementary strand. Biases with the same colour are summarized with one of the eight basic bias class labels (in the box at the bottom of the panel). **(b)** An example of a biased region in human promoter DNA, for colipase CLPS_1. The position in the promoter (downloaded from the EPD database (Schmid et al. 2004)) is indicated, along with the bias signature *{AC}*.

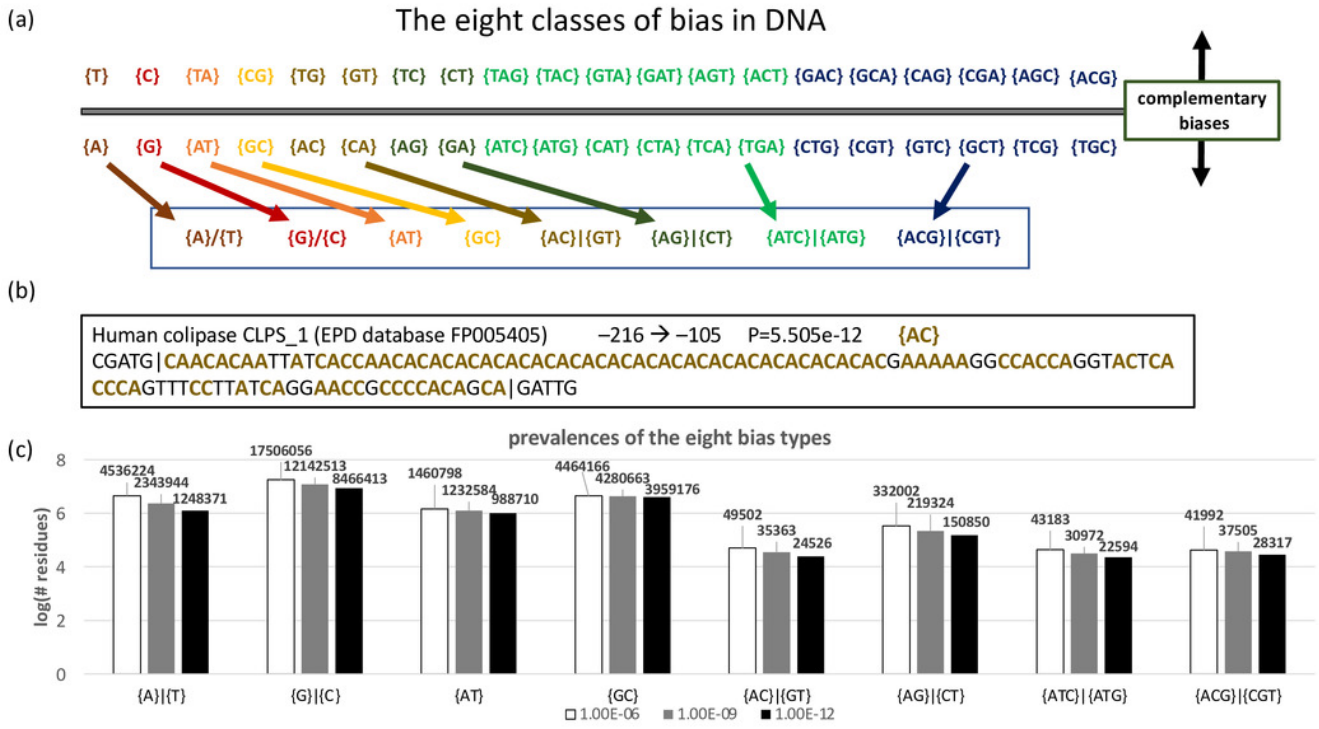
(a)
The eight classes of bias in DNA
{T} {C} {TA} {CG} {TG} {GT} {TC} {CT} {TAG} {TAC} {GTA} {GAT} {AGT} {ACT} {GAC} {GCA} {CAG} {CGA} {AGC} {ACG}
complementary biases
{A} {G} {AT} {GC} {AC} {CA} {AG} {GA} {ATC} {ATG} {CAT} {CTA} {TCA} {TGA} {CTG} {CGT} {GTC} {GCT} {TCG} {TGC}
{A}/{T} {G}/{C} {AT} {GC} {AC}|{GT} {AG}|{CT} {ATC}|{ATG} {ACG}|{CGT}
(b)
Human colipase CLPS_1 (EPD database FP005405) −216 → −105 P=5.505e-12 {AC}
CGATG|CAACACAATTATCACCAACACACACACACACACACACACACACACACACACACACACACACACGAAAAAGGCCACCAGGTACTCA
CCCAGTTTCCTTATCAGGAACCGCCCCACAGCA|GATTG
(c)
prevalences of the eight bias types
log(# residues)
0
2
4
6
8
4536224
2343944
1248371
17506056
12142513
8466413
1460798
1232584
988710
4464166
4280663
3959176
49502
35363
24526
332002
219324
150850
43183
30972
22594
41992
37505
28317
{A}|{T}
{G}|{C}
{AT}
{GC}
{AC}|{GT}
{AG}|{CT}
{ATC}|{ATG}
{ACG}|{CGT}
1.00E-06
1.00E-09
1.00E-12

Table 1(on next page)

Comparison of results for the precision options, using the yeast proteome as input

Table 1: Comparison of results for the precision options, using the yeast proteome as input*

	Number of single-residue CB regions			Number of multiple-residue CB regions		
P-value thresholds → **Precision option (–z)** **↓**	**P≤1e-03**	**P≤1e-06**	**P≤1e-09**	**P≤1e-03**	**P≤1e-06**	**P≤1e-09**
Fast (default)	32022	5781	2268	6336	4512	2744
Medium	36589	5792	2275	17117	6350	3395
Thorough	37738	5792	2276	27229	7675	3766

* UniProt reference proteome for S. cerevisiae 288C, downloaded January 2021.

Table 2(on next page)

Number of *S. cerevisiae* proteins with signal peptides that coincide with CB regions annotated by fLPS

Table 2: Number of *S. cerevisiae* proteins with signal peptides that coincide with CB regions annotated by fLPS*

P-value thresholds → **Precision option (–z)** ↓	**P≤1e-03**	**P≤1e-04**	**P≤1e-05**
Fast (default)	60/301 (19.9%)**	25/301 (8.3%)	15/301 (5.0%)
Medium	111/301 (36.9%)	51/301 (16.9%)	29/301 (9.6%)
Thorough	180/301 (59.8%)	90/301 (29.9%)	48/301 (15.9%)

* Annotations for signal peptides were taken from UniProt (301 in total).

** The numbers of signal peptides for which ≥50% of their residues overlap ≥50% of the residues of an individual fLPS-annotated CB region.