DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID baq031, doi:10.1093/database/baq031

Original article

LPS-annotate: complete annotation of compositionally biased regions in the protein knowledgebase

Djamel Harbi, Manish Kumar and Paul M. Harrison*

Department of Biology, McGill University, Stewart Biology Building, 1205 Dr. Penfield Ave., Montreal, QC, H3A 1B1, Canada

***Corresponding author:** Tel: +1 514 398 6420; Fax: +1 514 398 5069; Email: paul.harrison@mcgill.ca

Submitted 10 June 2010; Accepted 24 November 2010

Compositional bias (i.e. a skew in the composition of a biological sequence towards a subset of residue types) can occur at a wide variety of scales, from compositional biases of whole genomes, down to short regions in individual protein and gene–DNA sequences that are compositionally biased (CB regions). Such CB regions are made from a subset of residue types that are strewn along the length of the region in an irregular way. Here, we have developed the database server LPS-annotate, for the analysis of such CB regions, and protein disorder in protein sequences. The algorithm defines compositional bias through a thorough search for lowest-probability subsequences (LPSs) (i.e., the least likely sequence regions in terms of composition). Users can (i) initially annotate CB regions in input protein or nucleotide sequences of interest, and then (ii) query a database of greater than 1 500 000 pre-calculated protein-CB regions, for investigation of further functional hypotheses and inferences, about the specific CB regions that were discovered, and their protein disorder propensities. We demonstrate how a user can search for CB regions of similar compositional bias and protein disorder, with a worked example. We show that our annotations substantially augment the CB-region annotations that already exist in the UniProt database, with more comprehensive annotation of more complex CB regions. Our analysis indicates tens of thousands of CB regions that do not comprise globular domains or transmembrane domains, and that do not have a propensity to protein disorder, indicating a large cohort of protein-CB regions of biophysically uncharacterized types. This server and database is a conceptually novel addition to the workbench of tools now available to molecular biologists to generate hypotheses and inferences about the proteins that they are investigating. It can be accessed at http://libaio.biol.mcgill.ca/lps-annotate.html.

Database URL: http://libaio.biol.mcgill.ca/lps-annotate.html

Introduction

Development of tools for automated biological-sequence annotation is imperative, particularly since now greater than 1500 complete genomes have been sequenced and assembled. One important problem is the comprehensive annotation of compositionally biased (CB) regions in biological sequences. CB regions are sequence stretches with a large fraction of a small subset of residue types. If the CB regions are biased for multiple amino-acid residue types that are strewn along the sequence in an irregular way, the boundaries of these regions can be difficult to define. A well-known CB case arises in the yeast prions, which tend to contain CB regions made from glutamine and asparagine residue types (1). Other examples include the arginine-/serine-rich regions in some RNA-binding proteins (2), and the proline-rich domain of the transcriptional-complex protein Ssdp1, which is responsible for transactivation and is found in other diverse contexts (3).

A specific type of CB region is the 'intrinsically disordered' (ID) protein or domain. ID regions lack a globular 3D structure, and are unfolded in their native states (4,5). Work on disordered-region annotation has been extensive, with several algorithms being developed (5–10). ID proteins can function in signalling and regulation, and are associated with post-translational modifications, such as

Database, Vol. 2011, Article ID baq031, doi:10.1093/database/baq031

phosphorylation sites (5,11,12). The link between ID and CB regions, and simple repetitive regions has been demonstrated, with CB regions of a certain degree of disorder having distinct compartmentalizations and functional category tendencies (13–15).

Several algorithms have been derived previously to annotate CB regions, with the primary goal of 'masking' such regions before sequence alignment, to avoid incorrect inference of homology [e.g. SEG, (16); and CAST (17)]. To facilitate the automated annotation of all possible CB regions, we have developed a server, called LPS-annotate (LPS stands for Lowest Probability Subsequence; see the algorithm summary below for further details). This algorithm annotates CB regions in both protein and nucleotide sequences. Previously, we reported the development of this algorithm for the exhaustive assignment of CB regions (1,14,18). The chief novelty of this procedure is that CB regions of multiple amino acid residue types can be assigned thoroughly and completely, with clearly optimized boundaries.

Here, we report the development of a server and an on-line database of annotations that is based on the latest development of this algorithm. First, the server can be used to apply the LPS algorithm to annotation of both protein and nucleotide sequences. Second, after determining the biases in a sequence, a database of greater than 1500 000 pre-calculated CB regions and regions of predicted protein disorder (PPD) can be queried for regions of the same type and protein disorder content, for investigation of further functional hypotheses and inferences. In the database portion of the website, CB region annotations have been pre-calculated for the Uniprot/SWISSPROT database (19). CB annotations are cross-referenced with default SEG annotations of low-complexity regions (16), and also with predictions of disordered regions in proteins [made using DISOPRED2 (20)].

Methods

LPS algorithm

In brief, the LPS algorithm scans along the input sequence in a decreasing series of window sizes, the maximum (W_{max}) and minimum (W_{min}) of which are specified by the user. For each residue type x, and for the range of window sizes ($W_{min} \leq w \leq W_{max}$), the input sequence is searched for stretches that have compositional bias of the lowest probability (P_{min}):

$$P_{min} = \min[P\text{bias}(i, w)], \quad \forall i, x \tag{1}$$

where i is each possible start position for a window w in the sequence, spaced according to the user-specified parameter S (step size). The probability $P_{bias}(i, w)$ in Equation (1) is given by a binomial distribution:

$$P_{bias}(i, w) = [w!/[n!(w-n)!]] \bullet (f_x)^n \bullet (1-f_x)^{w-n} \tag{2}$$

where f_x is the proportion of amino-acid type x as given by the database amino-acid composition. The count for x is denoted n in the window w starting at position i. Sequence stretches with P_{min} are termed lowest-probability subsequences (LPSs).

To calculate biases derived from any number of residue types thoroughly for a given protein sequence, the following iterative process is performed. P_{min} values are calculated for any set of amino acids {*xyz*...}, by summing up the number of residues over the whole residue-type set. However, biases are only picked in preference over a previously calculated bias made by a smaller number of residue types, if their P_{min}-values are smaller. The set of residue types contributing to the bias (sorted in decreasing order of their original P_{min} values), is defined as the 'CB signature'. The iterative procedure is performed until convergence. Using this procedure, regions that comprise mild bias for multiple residue types can be detected as significantly biased. Further details of the algorithm are given on the help pages of the server [and in (14,18)].

Data analysed

The algorithm was run on the complete UniProt/SwissProt and UniProt/TrEMBL databases from July 2009 (19). The CB regions in the database are given a ranking, with the most biased (smallest *P*-value) being given a ranking of 1, and others given higher rankings.

Assignments of disordered regions in proteins were made using DISOPRED2 (20), with default settings. Of course, other DISOPRED program parameter settings are possible, but the disordered region predictions annotated here are just used as a guide, as a prelude to further detailed characterizations by the user for their proteins of interest. The fraction of the CB regions that are comprised of predicted protein disorder, was calculated from these assignments, and is displayed in the database entries. Also, the mean disorder propensity of each CB region was calculated by averaging the disorder propensity values for all residues in the CB region. Disorder propensity values ($P_{diso,\ X}$) were calculated from the DISPROT database of known disordered regions (21,22), for each amino-acid residue type X from the following formula:

$$P_{diso,\ X} = \frac{\left\lceil \begin{array}{l}[\text{Number of residues of type} \times \\ \text{in the disordered regions}] \\ /[\text{total number of residues} \\ \text{in the disordered regions}\end{array} \right\rceil}{\left\lceil \begin{array}{l}[\text{Number of residues in of type} \times \\ \text{in the whole database}] \\ /[\text{total number of residues} \\ \text{in the whole database}\end{array} \right\rceil}.$$

Assignment of globular domains was performed using blastp (*e*-value $\leq 1 \times 10^{-4}$) (23) comparisons to the ASTRALSCOP non-redundant database of protein domains made with a threshold of 40% sequence identity (24). Annotations of transmembrane domains were taken from the 'FT TRANSMEM' records in the UniProt/SwissProt database (19). Existing UniProt/SwissProt annotations of CB regions were taken from the 'FT COMPBIAS' records.

Use of the database

The database can be used in a two-step process:

(i) LPS-annotate server: assignment of CB regions in a query sequence (either protein or nucleotide sequences);
(ii) Database of pre-calculated CB annotations for UniProt: searching the database of pre-calculated CB regions for functional inferences (this is, of course, for protein sequences only).

LPS-Annotate server. The LPS-annotate server can annotate both protein and nucleotide sequences for CB regions. Users can paste the input sequence into the query box provided, and select values for W_{min}, W_{max} and S (step size). A screenshot of an example of the output of this server is illustrated, for a glutamine-/histidine-rich protein (Figure 1). A help page is provided, which explains the functioning of the server, including recommended values for W_{max} and W_{min}. Typically, if a smaller W_{max} is used, there are two effects on the CB region annotations: (i) longer CB regions are broken up into shorter stretches (of up to approximately the size of W_{max}); (ii) subsidiary mild biases that can only be detected with longer window sizes are not considered. Thus, it is generally advisable to use the largest W_{max} (= 500 residues length).

As shown in the Figure 1 example, for each CB region, the server output displays: (i) the protein name; (ii) the number of bias residues (i.e. those residue types that define the bias); (iii) the start and end points of the CB region; (iv) the CB region's binomial *P*-value; (v) the CB signature; (vi) the mean disorder propensity for the CB region (calculated as described in 'Methods' section); (vii) the CB region subsequences. Other fields in the output are explained in the downloadable Help page. A link to 'Download' the data is given at the bottom of the page.

Database of pre-calculated CB annotations for UniProt. We have supplied a database of CB-region annotations for proteins in the June 2009 version of the Uniprot/SWISSPROT protein database (19). These were made using the parameter settings for the LPS algorithm ($W_{min} = 25$, $W_{max} = 500$ and $S = 1$). The annotations in the database are cross-referenced with: (i) low-complexity regions identified with SEG (16) (run with default settings); (ii) predictions of disordered regions made with the program DISOPRED2 (run with default settings) (20). It is important to note here that the default settings for the SEG program are designed for sequence masking as a prelude to sequence alignment, not for the annotation of compositional biases, which is the purpose of the presently described algorithm, LPS-annotate.

The database can be searched in three ways: (i) with a Uniprot/SwissProt identifier; (ii) with a CB signature; or also (iii) with a sequence, through a BLAST search interface (23). The CB-signature search capability is particularly useful for finding regions of similar compositional bias and protein disorder content. Such similar regions may help infer functional linkages or hypotheses, in a sequence that was initially input into the LPS-annotate server. In the output for each database search, a list of CB regions in increasing

Results of search

Name	Number of Bias Residues	Start	End	P-value	Merged Status	Length of Bias	String	Propensity	Subsequence
sp|P07269|PHO2_YEAST	66	188	549	8.136422e-25	0	1	N	0.999	NSNNNYYFFD ICSITVGSWN RMKSGALQRR NFQSIKELRN LSPIKINNIM SNATDLMVLI SKKNSEINYF FSAMANNTKI LFRIFFPLSS VTNCSLTLET DDDIINSNNT SDKNNSNTNN DDDNDDNSNE DNDNSSEDKR NAKDNFGELK LTVTRSPTFA VYFLNNAPDE DPNLNNQWSI CDDFSEGRQV NDAFVGGSNI PHTLKGLQKS LRFMNSLILD YKSSNEILPT INTAIPTAAV PQQNIAPPFL NTNSSATDSN PNTNLEDSLF FDHDLLSSSI TNTNNGQGSN NGRQASKDDT LNLLDTTVNS NNNHNANNEE NHLAQEHLSN DADIVANPND HLLSLPTDSE LPNTPDFLKN TN
sp|P07269|PHO2_YEAST	26	22	60	6.648277e-21	0	2	QH	1.194	QHDQQQQQQQ QHDQQHNQQQ QPQPQPIQTQ NLEHDHDQH
sp|P07269|PHO2_YEAST	13	288	331	1.211776e-07	0	1	D	1.070	DDDIINSNNT SDKNNSNTNN DDDNDDNSNE DNDNSSEDKR NAKD
sp|P07269|PHO2_YEAST	8	8	60	5.128551e-05	2	1	H	1.111	HDFNTHFATD LDYLQHDQQQ QQQQQHDQQH NQQQQPQPQP IQTQNLEHDH DQH
sp|P07269|PHO2_YEAST	35	79	271	5.072881e-04	0	2	KI	0.982	KRTRAKGEAL DVLKRKFEIN PTPSLVERKK ISDLIGMPEK NVRIWFQNRR AKLRKKQHGS NKDTIPSSQS RDIANDYDRG STDNNLVTTT STSSIFHDED LTFFDRIPLN SNNNYYFFDI CSITVGSWNR MKSGALQRRN FQSIKELRNL SPIKINNIMS NATDLMVLIS KKNSEINYFF SAMANNTKIL FRI
sp|P07269|PHO2_YEAST	20	97	292	7.220772e-04	2	1	I	0.972	INPTPSLVER KKISDLIGMP EKNVRIWFQN RRAKLRKKQH GSNKDTIPSS QSRDIANDYD RGSTDNNLVT TTSTSSIFHD EDLTFFDRIP LNSNNNYYFF DICSITVGSW NRMKSGALQR RNFQSIKELR NLSPIKINNI MSNATDLMVL ISKKNSEINY FFSAMANNTK ILFRIFFPLS SVTNCSLTLE TDDDII

Figure 1. Screenshot of output of initial server portion of database. An example of the initial LPS-annotate program server output for the example PHO2, from budding yeast (P07269, PHO2_YEAST).

Database, Vol. 2011, Article ID baq031, doi:10.1093/database/baq031

order of binomial *P*-value is given (each with a link to the complete Database entries for each CB region) (Figure 2). Each CB-region name is a live link to the individual Database entry of the CB region (Figure 3). At the bottom of the page, a 'Download' link is provided, so that the user can download the list of similar CB regions (Figure 2).

An example of the individual database entry display for the glutamine/histidine-rich (QH-rich) region in PHO2 from budding yeast, is illustrated (Figure 3). PHO2 is a regulator in phosphate metabolism that contains a homeobox DNA-binding domain, and acts as a derepressor of PHO5, another central regulator. It binds to the upstream activator sequence of PHO5, and the promoters of TRP4, HIS4 and CYC1. The database entry contains the following useful information: (i) the subsequence identifier, a unique identifier for the subsequence in the UniProt/SwissProt sequence that is compositionally biased; (ii) sequence accession number; (iii) the initial bias used to build the CB region (in this case = 'Q'); (iv) the number of residues in the CB region defining the bias; (v) the start and end points of the CB region; (vi) the binomial *P*-value for the CB region; (vii) the rank of the CB region in the database; (viii) the CB signature (in this case = 'QH'); (ix) the mean protein disorder propensity (if >1.0, this indicates that the region on average has a propensity to protein disorder); (x) the proportion of the CB region that is disordered, according to the

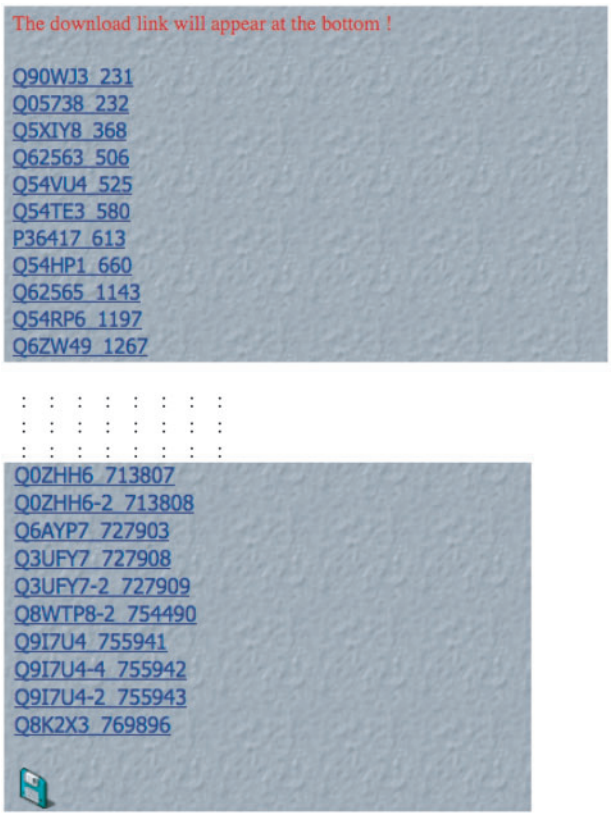

Figure 2. Screenshot of initial output after database search. An example of the initial LPS-annotate Database output for the search for bias type 'QH'. Each CB region is a live link in the depicted list. The download link for the data is at the bottom of the page.

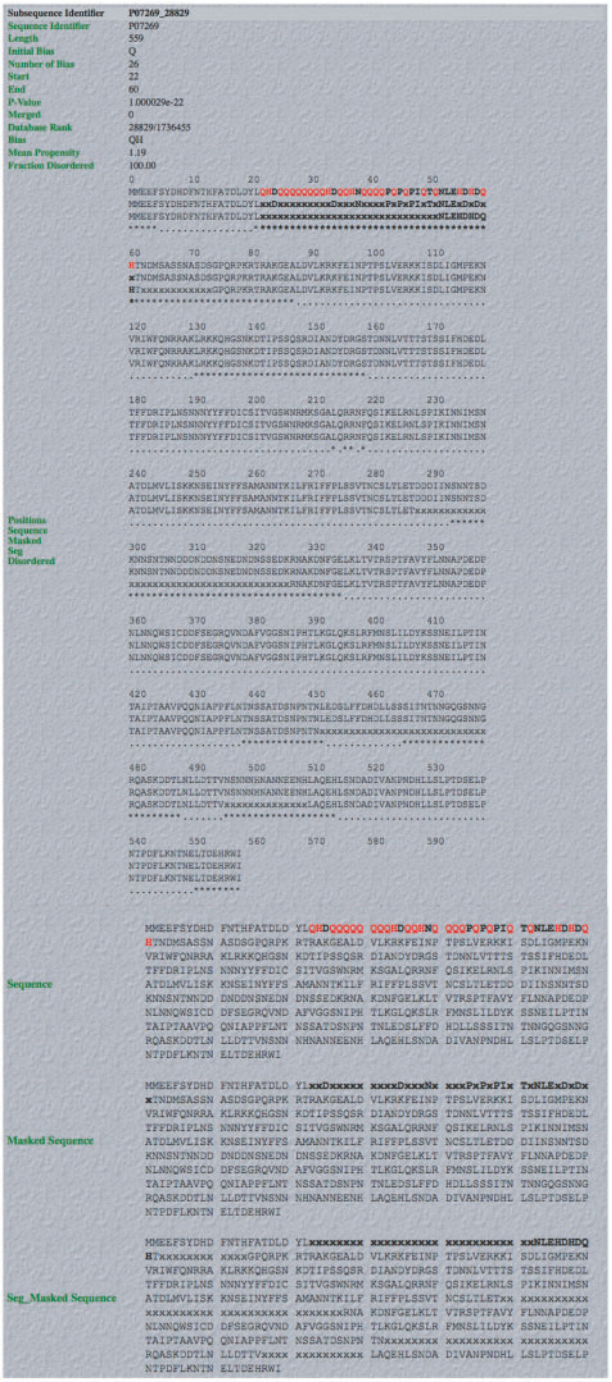

Figure 3. Screenshot of example of output from database. An example of a complete LPS-annotate Database entry (as described in the main text, and in the downloadable Database Help page). This is for the QH region from PHO2 of budding yeast.

disordered region assignments made with the DISOPRED2 algorithm (in this case it is 100.0%). Other database entry fields are described on the 'Help' page.

Below this list of information are displays of the sequence with the CB region in bold (and the bias-defining residues in red bold) (Figure 3). PPD (predicted using the DISOPRED 2 algorithm, see 'Methods' section) is indicated by asterisks (Figure 3).

The database and server for LPS-annotate is available at //libaio.biol.mcgill.ca/lps-annotate.html. A link is provided on the web page to download the complete database of annotations for both UniProt/SwissProt and UniProt/TrEMBL (July 2009 versions).

Using the LPS-annotate Database to search for similar CB regions in other proteins. The chief utility of the database is to find proteins with similar CB regions, to yield a list of proteins of use for further functional hypotheses and inferences. For example, take the sample sequence PHO2 from budding yeast. First, we can determine the CB regions in PHO2 using the LPS-annotate Program Server; the most obvious biased region in PHO2 is the 'QH'-rich region which is predicted to be 100% protein disorder by the program DISOPRED. Second, we can either: (i) click on the links given in the LPS-annotate Program Server output to obtain lists of similar biases in the LPS-annotate Database or (ii) type the biases of interest into the query box for the LPS-annotate Database Server, and proceed with downloading, from there. The output to download comprises a list of similar CB regions in other proteins, including the complete sequence of the CB regions within these other proteins. After download, this list of proteins can then be further examined bioinformatically by the user in a manner of his/her choosing, e.g. for shared globular protein domains elsewhere in the sequence, sequence motifs, cellular co-localizations and functional linkages [as specified, for example, by the Gene Ontology classification (25)].

Comparison of LPS-annotate Database to existing CB annotations in UniProt

We have substantially augmented the annotations of compositionally biased regions in the UniProt database, which are intentionally limited in the UniProt/SwissProt databases to a few, more specific cases, such as homopolymeric runs, with up to one or two short interruptions in the run (26). Here, we have generated more than 23 000 000 CB-region annotations for the UniProt/TrEMBL database, and more than 1 500 000 CB annotations for UniProt/SwissProt. Original CB-region annotations in UniProt number approximately 43 000 ('FT COMPBIAS' records); all of these are for the SwissProt portion of UniProt. We have compared these COMPBIAS feature annotations with our new annotations

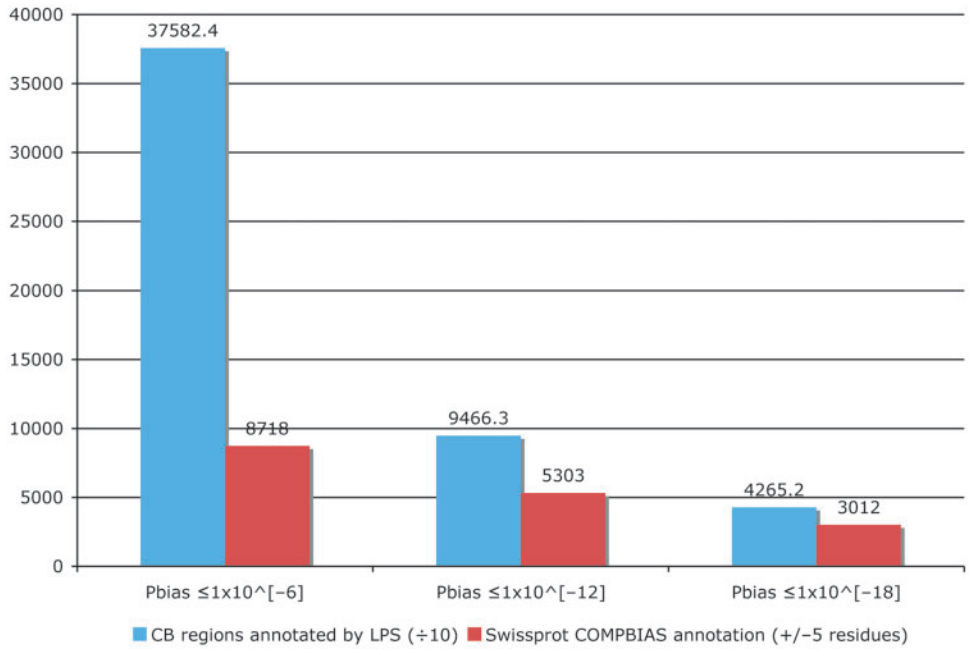

Figure 4. Column chart showing the augmentation of existing COMPBIAS annotations in UniProt, using the LPS algorithm. The blue column shows the number of CB regions annotated with the LPS-annotate algorithm, for three different *P*-value thresholds (10^{-6}; 10^{-12}; 10^{-18}). The red columns are the existing UniProt COMPBIAS records that overlap the new LPS-annotate annotations (±5 residues at either end of the regions). The UniProt COMPBIAS records are intentionally limited in the UniProt/SwissProt databases to a few, more specific cases, such as homopolymeric runs, with up to one or two short interruptions in the run (26).

Database, Vol. 2011, Article ID baq031, doi:10.1093/database/baq031

(Figure 4). For binomial $P \leq 10^{-6}$, the new CB annotations (blue column) overlap ~20% of the COMPBIAS records (to within five residues at either end point). A further breakdown of these overlapping CB annotations is given in some pie charts in Figure 5.

The CB-region annotations also do not have a simple correspondence with PPD. After removing CB region annotations corresponding to globular domains and transmembrane domains (Figure 6), there are a large number of CB regions without an overall tendency to protein disorder. For example, in Figure 6, for binomial $P \leq 10^{-6}$, ~21% of the CB regions (approximately 78 000 in number) have disorder propensity less than 1.0, indicating a cohort of CB regions that are potentially uncharacterized biophysical types, e.g. functional amyloids (27).

Conclusions

This server and database is a conceptually novel addition to the panoply of tools now available to molecular biologists

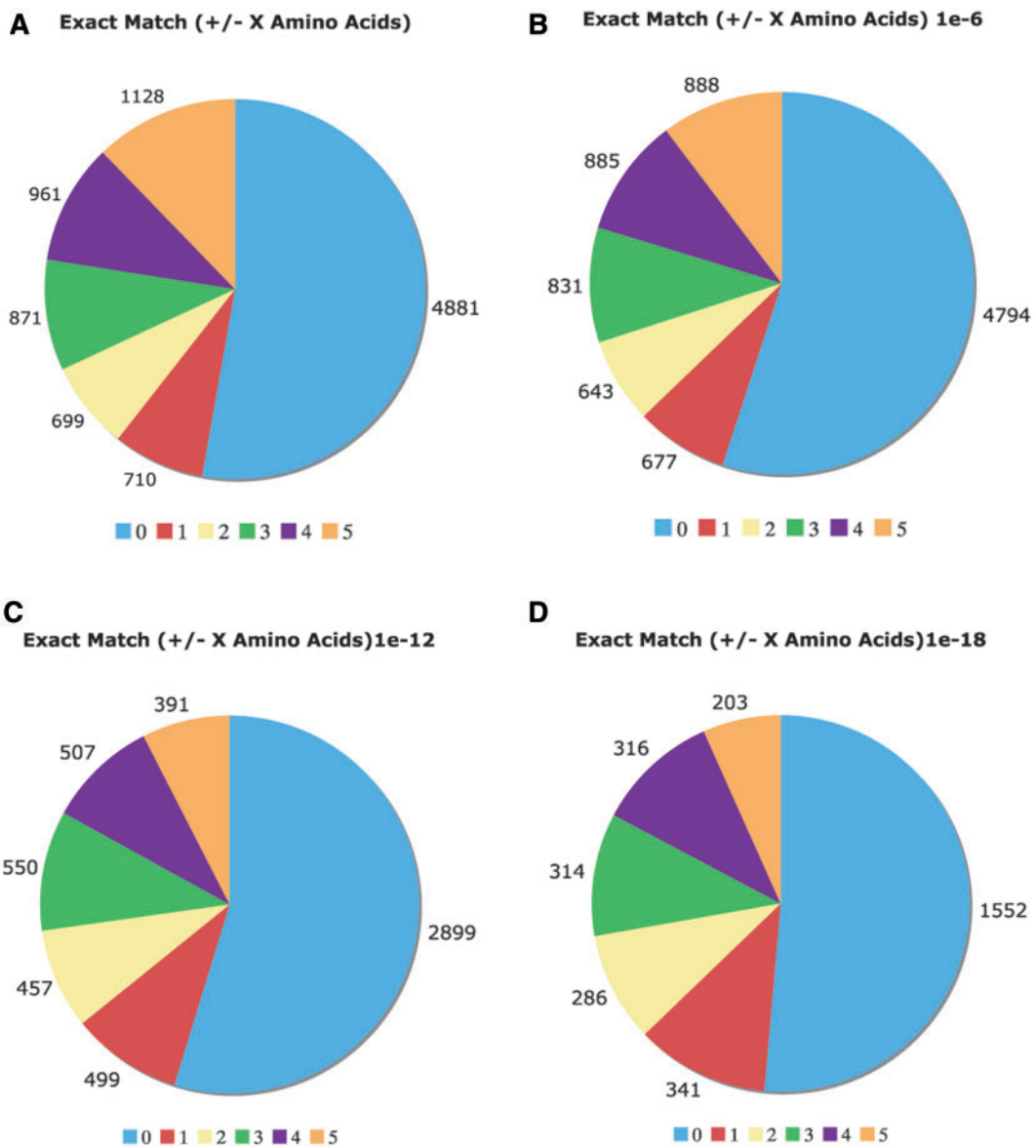

Figure 5. Comparison of the Uniprot COMPBIAS annotations with annotations by the LPS algorithm. Pie charts showing the detailed breakdown of how the new LPS-annotate CB annotations correspond with the UniProt COMPBIAS annotations, for four different *P*-value thresholds (10^{-4}, i.e. all of the LPS-annotate CB annotations; 10^{-6}; 10^{-12}; 10^{-18}). These are depicted in Figure parts **A**, **B**, **C** and **D** respectively. Annotations that are exactly matching are colored blue, those that are off by one at either end are colored red and so on. The UniProt COMPBIAS records are intentionally limited in the UniProt/SwissProt databases to a few, more specific cases, such as homopolymeric runs, with up to one or two short interruptions in the run (26).

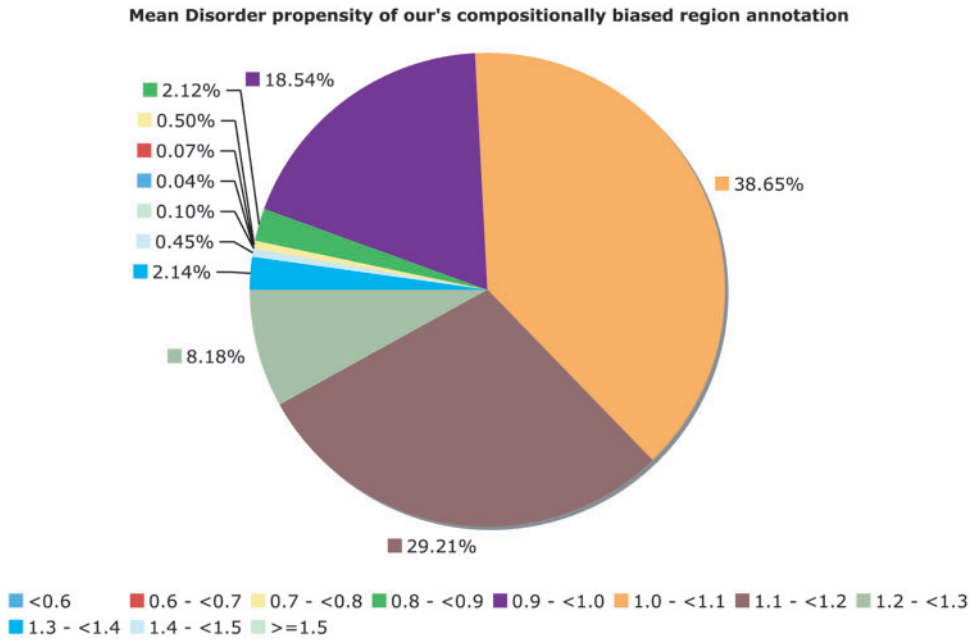

Figure 6. Mean disorder propensity of the CB regions. This is a pie chart for the mean disorder propensity of all CB regions with $P < 10^{-6}$, with any CB regions that correspond to globular or transmembrane domains removed. The mean disorder propensity is calculated as described in 'Methods' section.

to generate hypotheses and inferences about the proteins that they are investigating. Furthermore, large-scale analysis of cohorts of proteins with specific compositional biases and disorder propensities is made tractable by our analysis. The database of CB annotations is updatable at regular intervals.

Funding

This research was funded by the National Science & Engineering Research Council, Le Fonds québécois de la recherche sur la nature et les technologies, and McGill University. The open access publication charge is paid by Le Fonds québécois de la recherche sur la nature et les technologies.

Conflict of interest. None declared.

References

1. Harrison,L.B., Yu,Z., Stajich,J.E. *et al*. (2007) Evolution of budding yeast prion-determinant sequences across diverse fungi. *J. Mol. Biol.*, **368**, 273–282.
2. Long,J.C. and Caceres,J.F. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.*, **417**, 15–27.
3. Neduva,V. and Russell,R.B. (2007) Proline-rich regions in transcriptional complexes: heading in many directions. *Sci. STKE*, **2007**, pe1.
4. Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta.*, **1804**, 1231–1264.
5. Dunker,A.K., Silman,I., Uversky,V.N. and Sussman,J.L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.
6. He,B., Wang,K., Liu,Y. *et al*. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.
7. Dosztanyi,Z. and Tompa,P. (2008) Prediction of protein disorder. *Methods Mol. Biol.*, **426**, 103–115.
8. Dosztanyi,Z., Sandor,M., Tompa,P. *et al*. (2007) Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci.*, **8**, 161–171.
9. Bourhis,J.M., Canard,B. and Longhi,S. (2007) Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr. Protein Pept. Sci.*, **8**, 135–149.
10. Ferron,F., Longhi,S., Canard,B. *et al*. (2006) A practical overview of protein disorder prediction methods. *Proteins*, **65**, 1–14.
11. Gao,J., Agrawal,G.K., Thelen,J.J. *et al*. (2009) A new machine learning approach for protein phosphorylation site prediction in plants. *Lect. Notes Comput. Sci.*, **5462**, 18–29.
12. Iakoucheva,L.M., Radivojac,P., Brown,C.J. *et al*. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
13. Jorda,J., Xue,B., Uversky,V.N. *et al*. (2010) Protein tandem repeats - the more perfect, the less structured. *Febs J.*, **277**, 2673–2682.
14. Harrison,P.M. (2006) Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics*, **7**, 441.
15. Simon,M. and Hancock,J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.*, **10**, R59.
16. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.

17. Promponas,V.J., Enright,A.J., Tsoka,S. *et al*. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.
18. Harrison,P.M. 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–R46.
19. Apweiler,R., Bairoch,A., Wu,C. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
20. Ward,J.J., McGuffin,L.J., Bryson,K. *et al*. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.
21. Sickmeier,M., Hamilton,J.A., LeGall,T. *et al*. (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, **35**, D786–D793.
22. Vucetic,S., Obradovic,Z., Vacic,V. *et al*. (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
23. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. Chandonia,J.M., Hon,G., Walker,N.S. *et al*. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
25. Consortium, G.O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
26. UniProt. http://www.uniprot.org/manual/compbias (8 December 2010, date last accessed).
27. Fowler,D.M., Koulov,A.V., Balch,W.E. *et al*. (2007) Functional amyloid–from bacteria to humans. *Trends Biochem. Sci.*, **32**, 217–224.