Sequence analysis

A novel sensitive method for the detection of user-defined compositional bias in biological sequences

Igor B. Kuznetsov* and Seungwoo Hwang

Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, University at Albany, State University of New York, One Discovery Drive, Rensselaer, NY 12144, USA

Received on June 26, 2005; revised on December 23, 2005; accepted on February 7, 2006

ABSTRACT

Motivation: Most biological sequences contain compositionally biased segments in which one or more residue types are significantly overrepresented. The function and evolution of these segments are poorly understood. Usually, all types of compositionally biased segments are masked and ignored during sequence analysis. However, it has been shown for a number of proteins that biased segments that contain amino acids with similar chemical properties are involved in a variety of molecular functions and human diseases. A detailed large-scale analysis of the functional implications and evolutionary conservation of different compositionally biased segments requires a sensitive method capable of detecting user-specified types of compositional bias. Results: We present BIAS, a novel sensitive method for the detection of compositionally biased segments composed of a user-specified set of residue types. BIAS uses the discrete scan statistics that provides a highly accurate correction for multiple tests to compute analytical estimates of the significance of each compositionally biased segment. The method can take into account global compositional bias when computing analytical estimates of the significance of local clusters. BIAS is benchmarked against SEG, SAPS and CAST programs. We also use BIAS to show that groups of proteins with the same biological function are significantly associated with particular types of compositionally biased segments. Availability: The software is available at http://lcg.rit.albany.edu/bias/

Contact: ikuznetsov@albany.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

INTRODUCTION

Most protein and DNA sequences contain compositionally biased segments. In these segments one or more residue types are significantly overrepresented. Whether or not a given sequence segment has a compositional bias depends on the choice of background model used to define an unbiased sequence. In most biological applications residue positions in a protein or DNA sequence are modeled as a sequence of discrete events generated according to the random independence model. In this model a residue in sequence position \( j \), \( a_j \), is generated according to its probability, \( p(a_j) \), and this probability is independent of the rest of the positions (0-th order Markov model). Usually, \( p(a_j) \) is the same as the frequency of residue \( a_j \) computed using a large collection of protein or DNA sequences (some sequence database). The random independence model is used to assess the significance of database hits in the most popular and efficient sequence similarity search algorithms such as BLAST (Altschul et al., 1997). Therefore, compositionally biased segments that violate the assumptions of the random independence model may lead to incorrect estimates of statistical significance in sequence similarity searches and wrong functional and evolutionary inference.

Identification of compositionally biased segments is important not only in sequence similarity searches. DNA sequences contain a large fraction of diverse biased segments with poorly understood function. More than 50% of all proteins have been shown to contain biased segments (Wootton and Federhen, 1996). Some researchers believe that they are similar to junk DNA and are mostly determined by the compositional bias on the level of coding DNA sequence (Nishizawa and Nishizawa, 1998; Huntley and Golding, 2000; Singer and Hickey, 2000). Others suggest that biased segments in proteins are not just junk sequences and are involved in a variety of molecular functions: protein active sites, protein–DNA and protein–protein interactions (Karlin et al., 2003), transcription regulation (Brendel and Karlin, 1989), membrane transport (Kreil and Ouzounis, 2003), structural function (Karlin et al., 2003), peptides at protein termini (Berezovsky et al., 1999). In the case of a new protein that does not have detectable homologs, weak but statistically significant compositionally biased segments with certain properties can be one of few clues that the researcher can use to make a biologically meaningful guess about protein’s function and structure. Compositionally biased protein segments have also been implicated in a number of human diseases such as prion diseases, Huntington’s disease, cancer and others (Harrison and Gerstein, 2003; Karlin et al., 2002). It has been proposed that certain types of these segments are overrepresented in proteins involved in neurological disorders (Karlin and Burge, 1996). Unlike proteins, DNA sequences are composed only of four residue types and can be quite long. As a result, compositional bias in DNA is considerably harder to deal with than that in proteins.

There are two types of compositional bias: global and local. In the case of global bias, the entire protein sequence contains a large excess of particular residue types (an excess of hydrophilic residues, for instance) and becomes one long compositionally biased segment. In the case of local bias, most of the protein sequence conforms to the random independence model with only relatively

*To whom correspondence should be addressed.
small local clusters of over- or underrepresented residue types. Such locally biased segments may be the most likely candidates for functionally and/or structurally important sites. A number of methods that identify and mask compositionally biased segments by replacing them with lowercase letters or a string of ‘X’s in protein and ‘N’s in DNA sequences have been developed. The popular SEG algorithm uses a sliding window and entropy-based measure to detect regions with low-information complexity called low-complexity regions (Wootten and Federhen, 1996) that correspond to compositionally biased segments. Masking low-complexity regions with SEG significantly improves the specificity of sequence similarity search and it has become a de facto standard in protein BLAST searches. The SEG algorithm uses equal prior probabilities of residue types and segmentation threshold, chosen based on random sequences, and therefore provides no estimates of statistical significance of the detected low-complexity regions. It has been proposed that SEG reports too many low-complexity segments and introduces an artificial bias into the distribution of their length owing to a pre-set size of the window used to identify seed segments (Kreil and Ouzounis, 2003). CAST is another method for detecting compositionally biased segments (Promponas et al., 2000). It uses pairwise sequence alignment of query sequence with homopolymer sequences (20 homopolymer sequences in the case of proteins) and assigns significance based on the local alignment statistics. SIMPLE is a method for the identification of simple repeats in protein and DNA sequences that estimates statistical significance by using randomly generated sequences of the same length and composition as those of the test sequence (Alba et al., 2002). Neither of these methods is suitable for the identification of short biased segments composed of a user-specified subset of residue types with similar chemical properties.

The first statistical method designed for the identification of compositional bias by searching for statistically significant clusters of amino acid residues with similar chemical properties was proposed by Karlin et al. (1990) and implemented in program SAPS (Brendel et al., 1992). In this method, an amino acid sequence of length N is represented as a sequence of successes and failures in N independent Bernoulli trials. Successes correspond to residues that belong to a pre-selected group of similar amino acids (charged residues, for instance) and failures correspond to the rest of the 20 amino acids. Unusual clusters of amino acids from the group of interest are identified by counting the number of successes in a sliding window of fixed size. The significance of each window is estimated using the normal approximation to the binomial distribution. This approach has two major limitations. First, in order for the normal approximation to be valid window size must be sufficiently large. Second, this simple model does not provide a correction for multiple tests, a problem that arises from the use of all possible windows. In order to deal with multiple tests, SAPS uses three conservative p-value thresholds for proteins shorter than 750 residues, proteins with length 750–1500 residues, and proteins longer than 1500 residues. Because of these limitations, SAPS may miss weak but still significant and biologically informative signals in a form of small or low-density clusters. The program was designed neither for masking compositionally biased segments nor for the analysis of DNA sequences. Recently, a statistical method for finding compositionally biased regions in proteins similar to SAPS has been proposed (Harrison and Gerstein, 2003). This method also counts the number of successes in a sliding window and uses the exact binomial test to estimate statistical significance of each window. As in the case of SAPS, there is no explicit analytical correction for multiple testing. None of the above methods separates global compositional bias from the local one.

Here we present BIAS, a novel method and software for the identification of compositionally biased segments in protein and DNA sequences. BIAS automatically searches for unusual clusters of user-specified residue types and computes analytical estimates of the significance of each cluster. These estimates are based on scan statistics that allows one to detect even subtle local deviations from the random independence model. BIAS also distinguishes whether or not observed clusters are significant due to local or global compositional bias. Although the method is primarily intended for the analysis of proteins, it can be applied to DNA sequences as well. We use a number of biological examples to compare the performance of BIAS with that of SEG, SAPS and CAST programs. We also show that groups of proteins with the same biological function contain compositionally biased segments of the same type and that this association between type of function and type of compositional bias is highly non-random.

SYSTEMS AND METHODS

Overview

BIAS addresses the following problem. Given a sequence, S, of length N composed of residues generated using an alphabet, A, of size n according to the random independence model and a subalphabet, B, of m residues selected from A (m < n), find subsequences of S in which residues from B are overrepresented. These compositionally biased subsequences (segments) will correspond to clusters that have an unusually high density of residues from B and are unlikely to be observed by pure chance under the random independence model. In order to identify such clusters, S is represented as a sequence of successes and failures in N independent Bernoulli trials. Successes correspond to residues from B, failures correspond to those residues from A that are not included in B. Successes separated by less than d positions are merged into a single cluster. Statistical significance of each cluster is estimated using discrete scan statistic. Discrete scan statistics, denoted $S_{w}$, is the maximum number of successes observed within any w contiguous trials in a sequence of N Bernoulli trials. The tests based on scan statistics use a significance level that takes into account the sliding of the window to maximize the density of successes within the cluster. The unconditional discrete scan statistic is used in which the total number of successes in N trials is treated as a random variable. The probability of success is computed as the sum of probabilities of observing residue types from the subalphabet B.

Formal description of BIAS

Let $S = s_1, \ldots, s_N$ be a biological sequence, where each residue $s_i$ belongs to an alphabet $A = \{a_1, \ldots, a_n\}$. Let $B = \{b_1, \ldots, b_m\}$ be a subalphabet of A (m < n). Let $P(j)$ be the probability of occurrence of residue type j. Recode S into a sequence of N independent identically distributed binomial random variables $x_1, \ldots, x_N$, where $P(x_i = 1) = 1 - P(x_i = 0) = p$, according to the following rule:

$$x_i = \begin{cases} 1 & \text{(success)} \text{ if } s_i \in B \\ 0 & \text{(failure)} \text{ otherwise} \end{cases}$$

where $p = P(X_i = 1) = \sum_{j=1}^{m} P(b_j)$.

(1)

Use the linkage distance, d, to merge all positions between and including two successes $x_i$ and $x_j$ into a cluster, $C(i, j)$, if the following two conditions
are satisfied:

1. $C(i, j)$ cannot be extended—subsequences $x_i - d, \ldots, x_i - 1$ and $x_j + 1, \ldots, x_j + d (i - d \geq 1$ and $j + d \leq N$) consist only of failures or $i = 1$ or $j = N$ (positions at sequence termini).

2. Any two consecutive successes $x_i$ and $x_j$ from the cluster are separated by less than $d$ positions: for $i \leq f < g \leq i, g - f \leq d$.

For a given cluster $C(i, j)$ compute the number of successes, $k$, and cluster size, $w = j - i + 1$. Estimate the significance of the cluster by computing the probability of observing the value of unconditional discrete scan statistic, $S_w$, greater or equal to $k$ in a sequence of $N$ trials with the probability of success $p$, $P(S_w \geq k | N, p)$. This probability can be computed analytically using the following highly accurate approximation (Glaz et al., 2001):

$$P(S_w \geq k | N, p) = 1 - Q_1 \left( \frac{Q_1(N-w-v)}{\binom{N-w-v}{k}} \right),$$

where $v = 2$ if $N$ is a multiple of $w$ and $v = 1$ otherwise. For $2 < k < N$, we have

$$Q_1 = (F_b(k-1;w,p))^2 - (k-1) \cdot b(k;w,p) \cdot F_b(k-2;w,p)$$
$$+ w \cdot p \cdot b(k;w,p) \cdot F_b(k-3;w-1,p)$$

$$Q_2 = (F_b(k-1;w,p))^2 - A_1 + A_2 + A_3 - A_4$$

$$b(k;w,p) = \sum_{i=0}^{k} b(i;w,p) \text{ if } k = 0, 1, \ldots, w$$
$$0 \text{ if } k < 0$$

$$A_1 = 2 \cdot b(k;w,p) \cdot F_b(k-1;w,p) \cdot \left\{ (k-1) \cdot F_b(k-2;w,p) \right\} - w \cdot p \cdot b(k;w,p) \cdot F_b(k-3;w-1,p)$$

$$A_2 = 0.5 \cdot (b(k;w,p))^2 \cdot \left\{ (k-1) \cdot (k-2) \cdot F_b(k-3;w,p) \right\} - 2 \cdot (k-2) \cdot w \cdot p \cdot F_b(k-4;w-1,p)$$
$$+ w \cdot p \cdot F_b(k-3;w-1,p)$$

$$A_3 = \sum_{i=1}^{k-1} b(2k - r;w,p) \cdot (F_b(r-1;w,p))^2$$

$$A_4 = \sum_{r=2}^{k-1} b(2k - r;w,p) \cdot b(r;w,p) \cdot \left\{ (r-1) \cdot F_b(r-2;w,p) \right\} - w \cdot p \cdot b(r;w,p) \cdot F_b(r-3;w-1,p).$$

The significance of global compositional bias of sequence $S$ is estimated using the binomial distribution. Let $h$ be the total number of occurrences of residues from $B$ in $S$. Then, for an excess of residues from $B$ we estimate $P(X \geq h | N, p)$ for a lack of residues from $B$ we estimate $P(X < h | N, p)$ by the following equations:

$$P(X \geq h | N, p) = \sum_{i=h}^{N} b(i;N,p)$$
$$P(X < h | N, p) = \sum_{i=0}^{h-1} b(i;N,p).$$

![Flowchart](https://example.com/flowchart.png)

**Input amino acid sequence of length 20 ($N=20$)**

- Recode the sequence by assigning success, 1, to negatively charged residues (D or E) and failure, 0, to the remaining 18 amino acid
- Sequence of 20 trials with 4 successes and 16 failures. Probability of success: $p_S = P(D) + P(E) = 0.05$ by default.

**Cluster of 4 successes in a window of size 5 ($S_5 = 4$)**

- Estimate significance of the observed cluster of 4 successes in a window of size 5, $P(S_5 \geq 4 | N=20, p_S)$ using Eq. 2.
- Report sequence segments that correspond to clusters of successes with $p$-value below threshold (0.05 by default).

**Output masked sequence**

**ACTLVDDVEDSTLLMATHVK**

**Fig. 1.** Flowchart that shows an application of BIAS to search for clusters of negatively charged residues in a protein sequence.

**Implementation**

The method is implemented as a standard ANSI C++ program. The program (PBIAS for proteins and NBIAS for DNA) has a command-line interface, reads input sequences in standard FASTA format and is suitable for high-throughput sequence analysis pipelines. Source code and pre-compiled binaries for Windows and Linux are freely available for non-commercial use at http://leg.rit.albany.edu/bias/. A Perl script, MPBIAS.PL, that implements a CAST-like (Promponas et al., 2000) procedure for masking of low-complexity segments in a protein sequence is also provided. This script searches for subsequences in which one of the 20 amino acid types is significantly overrepresented and then replaces all positions in such subsequences with the ‘X’ character.

Probability of each of the 20 amino acid types is estimated using its frequency in a non-redundant protein database. Two sets of frequencies are used: SPROT50, derived from the SWISS-PROT database (Bairoch and Apweiler, 2000) and PDB50, derived from the Protein Databank proteins (Berman et al., 2000). Sequences in both datasets were clustered at 50% sequence identity, meaning that all sequence pairs within the same dataset have identity <50%. Clustering was performed using the CD-HIT program (Li et al., 2002). PDB50 frequencies serve as the background model derived.
from mostly globular proteins, whereas SPROT50 frequencies serve as the background model derived from all proteins in the protein universe. In order to remove the effect of global compositional bias, significance of each cluster is also estimated using sequence-derived residue frequencies (default option for DNA sequences). The program also has options for estimating the significance of local compositional bias by using random sequences.

**DISCUSSION**

In this section we present the application of PBIAS to a number of well-annotated proteins that contain domains with a different type of chemical bias that determines specific properties of these domains and to a set of compositionally biased proteins from the Plasmodium genome. The performance of PBIAS is compared with that of SEG (Wootton and Federhen, 1996), SAPS (Brendel et al., 1992) and CAST (Promponas et al., 2000). For a reasonably fair comparison, all programs were used with default arguments (linkage distance of 4, SPROT50 frequencies and $p$-value threshold of 0.05 for PBIAS). These arguments were shown to perform well on a wide variety of compositionally biased proteins. Depending on the nature of each protein site, PBIAS was used with a subalphabet that contains amino acid types with similar chemical properties suitable for the identification of this particular site: hydrophobic, aliphatic, charged, etc. SAPS was used via web interface available at http://www.ch.embnet.org/software/SAPS_form.html. CAST was used via web interface available at http://biophysics.biol.uoa.gr/cgi-bin/CAST/cast.cgi. We also show that NBIAS can be successfully used to mask a cDNA sequence coding for a compositionally biased protein domain. Finally, we use BIAS to demonstrate a highly significant non-random association between type of function and the peptide as a transmembrane region. SAPS also reports the same peptide as a hydrophobic region, 240–253 versus 240–252 in PBIAS. PBIAS accurately identifies signal peptide residues 232–253 when linkage distance is increased from the default value of 4 to 5.

(1) PBIAS is the only program that accurately identifies the N-terminal signal peptide as an unusually dense cluster of hydrophobic residues. PBIAS, SEG and SAPS identify the C-terminal signal peptide. In this case, SAPS is most accurate with default parameters (residues 232–253), but it reports the peptide as a transmembrane region. SPROT50 also reports the same peptide as a hydrophobic region, 240–253 versus 240–252 in PBIAS. PBIAS accurately identifies signal peptide residues 232–252 when linkage distance is increased from the default value of 4 to 5.

(2) PBIAS accurately identifies most of the unstructured N-terminal domain as a highly significant long cluster of residues with high propensity for irregular (coil) conformation. CAST also identifies a significant portion of this domain. Both SEG and SAPS identify only about half of this domain. Unlike PBIAS, SEG and SAPS do not provide any insights into the properties of the residues in this domain.

(3) PBIAS and SEG identify the flexible C-terminal end of helix B. PBIAS identifies it as a highly significant cluster of Thr695 residues and is more accurate than SEG. Neither SAPS nor CAST report this segment.

It should be noted that all segments identified in PrP by PBIAS are statistically significant when the global compositional bias of the protein is accounted for. In other words, they all represent true cases of local compositional bias. This is especially interesting in the case of the unstructured N-terminal domain. Although PrP has a significant excess of residues with high propensity for coil protein has been extensively characterized experimentally. It consists of a flexible unstructured N-terminal domain thought to be involved in binding copper and a structured mostly helical C-terminal domain. It has been shown that upon formation of PrP dimer the switch region located in the C-terminus of helix B unwinds (Knaus et al., 2001). PrP also contains N- and C-terminal signal peptides which are removed during post-translational modification (Lehmann et al., 1999). The results of application of PBIAS, SEG, SAPS and CAST to human prion protein are shown in Table 1 and briefly summarized below.

**Table 1.** Comparison of compositionally biased regions detected by PBIAS, SEG, SAPS and CAST in human prion protein, PDB id: 1QLX, Swiss-Prot id: P04156

<table>
<thead>
<tr>
<th>Functionally/structurally annotated region</th>
<th>PBIAS regions</th>
<th>SEG regions</th>
<th>SAPS regions</th>
<th>CAST regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal signal peptide, residues 1–22</td>
<td>1–22, $p = 2 \times 10^{-2}$</td>
<td>N/D</td>
<td>N/D</td>
<td>N/D</td>
</tr>
<tr>
<td>Flexible unstructured N-terminal domain, residues 23–124</td>
<td>26–108, $p = 2 \times 10^{-7}$</td>
<td>50–94</td>
<td>54–94 (identified as simple tandem repeat)</td>
<td>29–94 (G-rich)</td>
</tr>
<tr>
<td>Fragment of helix B that unwinds upon prion dimerization, residues 189–195</td>
<td>188–199, $p = 6 \times 10^{-4}$</td>
<td>188–201</td>
<td>N/D</td>
<td>N/D</td>
</tr>
</tbody>
</table>

N/D, not detected.

$^a$PBIAS detects this region only if sequence-derived amino acid frequencies are used.

$^b$Segment remains significant if global compositional bias is accounted for.

**Prion protein**

The prion protein, PrP, is a rare infectious agent that causes transmissible spongiform encephalopathy. PrP can undergo a dramatic transition from the normal mostly helical conformation to the pathogenic conformation rich in beta-sheet (Prusiner, 1998). Prion
conformation [Equation (11), \(p = 3 \times 10^{-3}\)], the cluster of these residues corresponding to the N-terminal domain is significant \(\langle p = 5 \times 10^{-4} \rangle\) even when sequence-derived frequencies are used.

Ser-tRNA synthetase

The N-terminal domain of Ser-tRNA synthetase is a non-globular long helical bundle that interacts with cognate tRNA (Fig. 2) (Fujinaga et al., 1993). PBIAS accurately identifies this domain as a highly significant \(\langle p = 4 \times 10^{-7}\rangle\) cluster of strong helix formers (Table 2). Despite a large excess of strong helix formers [Equation (11), \(p = 7 \times 10^{-6}\)], this cluster remains significant \(\langle p = 8 \times 10^{-3}\rangle\) when global bias is accounted for. SEG identifies two short low-complexity segments that cover only \(\approx 35\%\) of this domain. CAST identifies about two-third of this domain. SAPS does not report any unusual segments.

Collagen

Collagen is an extracellular structural protein with three long non-globular triple-helical domains rich in amino acids Glycine and Proline that have unusual backbone conformational preferences. These triple-helical domains are involved in the formation of coiled-coil structure (Beck and Brodsky, 1998). PBIAS is most accurate at identifying these domains as highly significant clusters of Glycine and Proline (Table 2). These clusters remain significant when global bias is accounted for. SEG identifies many short low-complexity regions in these domains. SAPS identifies the entire 275–897 segment as a highly repetitive region. CAST identifies 269–899 segment as Gly- and Pro-rich.

Cytochrome C biogenesis protein

Cytochrome C biogenesis protein is involved in the heme delivery pathway for cytochrome c maturation (Stevens et al., 2004). This protein contains a transmembrane domain (residues 67–88). SAPS is most accurate at identifying this domain with default parameters. PBIAS with default arguments and subalphabet of aliphatic residues (lipid-soluble residues) identifies a slightly longer segment that includes this domain. If linkage distance is changed from the default value of 4 to 3, PBIAS identifies the same domain as SAPS (Table 2). SEG identifies only a part of the transmembrane domain. Both PBIAS and SEG find another compositionally biased segment in the C-terminus. CAST does not report any segments.

Ribosomal protein L32E

Ribosomal protein L32E has two disordered regions, long (residues 1–94) and short (residues 237–340) (Klein et al., 2001). PBIAS is most accurate at identifying a long region as an unusual cluster of residues with high propensity for disordered conformation. Both SEG and SAPS identify only a small fraction of this region, SAPS identifying it as a negatively charged region (Table 2). CAST reports the 2–113 segment as Glu-rich.

C-Myc I protein

C-Myc I protein has an acidic domain (residues 226–245) and a leucine zipper domain (residues 387–415) (Viriz et al., 1989). PBIAS is most accurate at identifying these domains as clusters of charged residues (Table 2), whereas SEG reports too many low-complexity regions. Both CAST and SAPS identify the acidic domain but miss the leucine zipper.

Masking cDNA sequence of prion protein

NBIAS was used with default arguments (linkage distance 5, sequence-derived frequencies, \(p\)-value threshold of 0.05) and subalphabet GC to mask the cDNA of human prion protein. This procedure resulted in masking almost entire N-terminal part (codons 2–105) which is significantly compositionally biased in the protein sequence. SEG only masked codons 96–105. Application of NBIAS to the cDNA sequence of chicken PrP that has very different nucleotide composition and shares only 35% sequence identity with human PrP gives similar results (see Supplementary information).

Compositionally biased segments in PDB and SWISS-PROT proteins

It has been shown before using SEG that low-complexity segments are underrepresented in globular proteins available in the Protein Databank (PDB) (Huntley and Golding, 2002). Application of PBIAS to proteins from the PDB and SWISS-PROT confirms this observation. As one can see from Figure 3, the percentage of proteins with at least one compositionally biased segment that has \(p\)-value \(< 10^{-3}\) is significantly lower in PDB proteins for all tested subalphabets. The largest difference is observed for hydrophobic and hydrophilic segments that are underrepresented in PDB by more than one order of magnitude. This is consistent with the observation that sequences lacking periodicity of hydrophobic/hydrophilic residues cannot form compact globular structure (Dill, 1999; Silverman, 2005) and therefore are underrepresented in mostly globular PDB proteins.

Analysis of low-complexity segments in Plasmodium falciparum proteins

In order to compare the performance of PBIAS with that of SEG and CAST on a large sequence set we used 210 proteins from chromosome 2 of P. falciparum (Gardner et al., 1998). We chose this dataset because it represents a set of highly compositionally
biased eukaryotic proteins that was extensively used before to compare CAST and SEG methods (Promponas et al., 2000). Each of the three programs was used with the default arguments. The distribution of the number of low-complexity segments detected in this dataset by PBIAS, SEG and CAST is shown in Figure 4. We find that performance of PBIAS is more similar to that of SEG than CAST. For instance, PBIAS identifies at least one low-complexity segment in 93% of the sequences, SEG in 90%, whereas CAST only in 74%. Both PBIAS and SEG also report significantly higher number of sequences that contain five or more low-complexity segments than CAST.

We also studied how PBIAS masking of low-complexity segments performs in database searching compared with masking with SEG and CAST. We used each of the 210 P.falciparum sequences masked by PBIAS, SEG and CAST as a query sequence in BLASTP (Altschul et al., 1997) sequence similarity search. BLASTP was run against the NCBI NR database with the default arguments and masking option turned off. For each query
sequence, \(i\), we computed four overlap indices normalized between 0 and 100%:

\[
O_S^P(i, E) = \frac{N_{SEG\cdot PBIAS}(i, E)}{N_{SEG}(i, E)}
\]

\[
O_P^P(i, E) = \frac{N_{SEG\cdot PBIAS}(i, E)}{N_{PBIAS}(i, E)}
\]

\[
O_C^P(i, E) = \frac{N_{CAST\cdot PBIAS}(i, E)}{N_{CAST}(i, E)}
\]

\[
O_P^O(i, E) = \frac{N_{CAST\cdot PBIAS}(i, E)}{N_{PBIAS}(i, E)}
\]

where \(N_X(i, E)\) is the total number of hits with BLASTP \(-\) value below threshold \(E\) returned for query sequence \(i\) masked with method \(X\). \(N_{SEG\cdot Y}(i, E)\) is the number of hits with \(-\) value below threshold \(E\) that appear in both hit list for \(i\) masked with method \(X\) and in hit list for \(i\) masked with method \(Y\) (size of the overlap between the two lists). \(X, Y \in \{PBIAS, CAST, SEG\}\). If the value of a given overlap index, \(O_X^P\), is 100% it means that all hits in list \(X\) are also included in list \(Y\). Small values of the overlap index mean that list \(X\) contains many hits that are not included in list \(Y\). We used two \(-\) value thresholds, \(10^{-6}\) and \(10^{-30}\).

The results of BLASTP benchmarking indicate that when PBIAS is compared with SEG, identical hit lists are returned for 30% of the sequences if \(-\) value is set to \(10^{-6}\) and for 41.4\% of the sequences if \(-\) value is set to \(10^{-30}\). When PBIAS is compared with CAST, identical hit lists are returned for 31.4 and 39.5\% of the sequences, respectively. A more detailed analysis of the overlap indices (Table 3) shows that both SEG- and CAST-masked query sequences tend to return more hits than the PBIAS-masked ones (as indicated by low average \(O_C\) and \(O_S\) overlap indices ranging from 67.6 to 76.0\%). Almost all hits returned for PBIAS-masked sequences are included in SEG and CAST hit lists (as indicated by high average \(O_P\) overlap indices ranging from 95.2 to 99.7\%).

### Table 3. The average and standard deviation of the overlap between hit lists returned for comparisons PBIAS versus SEG and PBIAS versus CAST performed using the database searching benchmark

<table>
<thead>
<tr>
<th></th>
<th>(O_S^P) (Eq. 12)</th>
<th>(O_P^P) (Eq. 12)</th>
<th>(O_C^P) (Eq. 13)</th>
<th>(O_P^O) (Eq. 13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTP (-)value = (10^{-6})</td>
<td>95.2% ± 13.0%</td>
<td>74.1% ± 36.0%</td>
<td>98.0% ± 10.0%</td>
<td>67.6% ± 33.1%</td>
</tr>
<tr>
<td>BLASTP (-)value = (10^{-30})</td>
<td>96.8% ± 12.2%</td>
<td>72.0% ± 37.5%</td>
<td>99.7% ± 2.3%</td>
<td>67.6% ± 37.9%</td>
</tr>
</tbody>
</table>

See text for details.

Identify a significant association between biased segments from a subalphabet \(C\) and a particular GO term \(i\):

1. Use PBIAS with default arguments and a given subalphabet, \(C\), to search each protein from SPROT50 dataset for biased segments. For GO term \(i\) count the number of proteins, \(n(i)\), that contain at least one biased segment with \(-\) value below the threshold of \(10^{-3}\) and are linked to this term.

2. Count the number of proteins in SPROT50, \(M\), that contain at least one biased segment from subalphabet \(C\). Randomly sample, without replacement, \(M\) proteins from SPROT50. Count the number of proteins in this random sample associated with GO term \(i\), \(r(i)\). Repeat the sampling \(N\) times.

3. Use random samples to compute for each \(r(i)\) its average, \(E(i)\), and standard deviation, \(S(i)\). Use these to compute the Z-score for GO term \(i\), \(z(i)\):

\[
z(i) = \frac{n(i) - E(i)}{S(i)}
\]

where

\[
E(i) = \frac{\sum_{j=1}^N r(j)}{N}
\]

and

\[
S(i) = \sqrt{\frac{\sum_{j=1}^N (r(j) - E(i))^2}{N - 1}}
\]

Random sampling was performed \(10^5\) times (\(N = 10^5\), a number for which the Z-score converges up to one decimal digit).

We used two non-overlapping subalphabets, RKENDSTQ (hydrophilic amino acids) and LVIFM (major hydrophobic amino acids), and a subalphabet that partially overlaps with the hydrophilic one, GPTSDN (amino acids with high propensity for coil conformation). Figure 5 shows all GO terms from ‘Molecular function’ ontology that have the absolute value of \(z(i)\) greater than 6.0 for at least one of the three subalphabets. Since we cannot assume that all distributions are normal, the threshold Z-score of 6.0 was chosen to ensure that the observed associations are significant even for skewed distributions. Inspection of Figure 5 leads to the following conclusions:

1. There are groups of GO terms associated with only one type of biased segments (hydrophobic, hydrophilic or coil), despite the fact that hydrophilic and coil subalphabets overlap. Hydrophobic clusters are associated with receptor activity, voltage-gated channel activity, transporter activity. Hydrophilic clusters are associated with nuclease activity,
transcription, splicing and translation. Clusters of coil residues are associated with direct binding to nucleic acids, protein binding, structural constituent of cuticle.

(2) Presence of hydrophobic clusters is negatively correlated with the presence of hydrophilic clusters. There are, however, three exceptions: GO:4872 (receptor activity), GO:5245 (voltage-gated calcium channel activity) and GO:5486 (t-SNARE activity), all being integral membrane proteins.

Very similar trends are observed in the other two ontologies, ‘Cellular component’ and ‘Biological process’. However, the number of significant associations in these two ontologies is very large and cannot be presented in a compact form suitable for the limited journal space. Detailed results for a variety of subalphabets and all three ontologies will be presented elsewhere. The primary point from Figure 5 is that we do observe that groups of proteins with the same biological function contain the same type of compositionally biased segments and that this association between type of function and type of compositional bias is highly non-random. This suggests that compositionally biased segments are not just ‘junk sequences’ and can be used to construct potential function-specific de novo protein signatures.

CONCLUSIONS

We presented BIAS, a method for the identification of compositionally biased segments composed of a user-specified set of residue types. Main features of the method are as follows.

(1) BIAS finds statistically significant clusters of user-specified residue types that are unlikely to be observed under the random independence model. Significance is estimated analytically using scan statistics that provides a highly accurate correction for multiple tests.

(2) BIAS provides an exact estimate of global compositional bias of the test sequence and can take into account global compositional bias when computing analytical estimates of the significance of local clusters.

(3) Both the main advantage and main shortcoming of the proposed method is that it requires the user to supply a subalphabet of residue types used to search for compositional bias. The advantage is that BIAS can be used to search for compositionally biased segments composed of amino acids with similar chemical properties such as charge, hydrophobicity, size, etc. The disadvantage is that BIAS will ignore biased regions composed of residue types that are not included into the user-supplied subalphabet.

The application of BIAS to a number of test proteins has shown that it is suitable for de novo characterization of functionally important sites. The application of BIAS to study proteins from the SWISS-PROT database has also shown that groups of proteins with the same biological function contain the same type of compositionally biased segments and that this association between type of function and type of compositional bias is non-random.

Fig. 5. GO functional categories from ‘Molecular function’ ontology that are significantly associated with clusters of residues with particular chemical properties. The following subalphabets were used: RKNEQDT, major hydrophilic amino acids; LVFM, major hydrophobic amino acids; GPDNST, amino acids with high preference for coil conformation. The Z-score given by the height of the bar quantifies the strength of the association between a particular GO functional category given by GO id (x-axis) and a particular type of residue clusters. The type of residue cluster is given by the shading pattern (white, black or striped bars). A complete description of each GO term is given in Supplementary information.
ACKNOWLEDGEMENTS

The authors thank Dr Yaroslava Ruzankina for help and two anonymous reviewers for their valuable comments.

Conflict of Interest: none declared.

REFERENCES

Li,W. et al. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics, 8, 77–82.