Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Sequence Features of Compositionally Biased Regions in Three Dimensional Protein Structures

Stella Tamana, Ioannis Kirmitzoglou, and Vasilis J. Promponas
Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus,
P.O. Box 20537, CY 1678, Nicosia, Cyprus
Email: {stella.tamana, ioannis.kirmitzoglou, vprobon}@ucy.ac.cy

***Abstract*—A considerable research effort has already been put on the identification (and consequently filtering) of local segments of "unusual" composition (Compositionally Biased or Low Complexity Regions; CBRs or LCRs) in protein sequences. This interest was mainly initiated due to the fact that CBR existence is known to create artifacts (i.e. biologically irrelevant hits) in sequence database search methods. Even though no general biological significance has been demonstrated for CBRs so far, they are often associated with the lack of regular structure. However, application of commonly used methods for CBR detection illustrates that instances of CBRs can be found in proteins with experimentally determined three dimensional structures. In this work, we highlight sequential properties of CBRs detected by two of the most widely used CBR detection algorithms in carefully compiled datasets of proteins with experimentally determined structures. Our goal is to shed light on the properties of CBR sequences, with the future prospect of elucidating their relation to protein three dimensional structure.**

I. Introduction

For practical purposes, protein sequences are often considered to be random sequences composed by the twenty standard amino acid types [1],[2]. However, there exist protein sequences containing local segments of "unusual" amino acid composition [3], showing preference in the appearance of a subset of amino acid residues (typically one or a few residue types). Depending on the formalism used to lead to their detection these local segments are called Low Complexity Regions (LCRs) [4] or Compositionally Biased Regions (CBRs)[5][1]. Another working definition for CBRs is 'Simple Sequences' [6], and Huntley and Golding (2002) described this subset as the perfect repeats of a single amino acid, observing their excess in eukaryotes but not in prokaryotes [7]. However, Simple Sequences are considered as a subset of the Low Complexity Sequences ([6], [8]) and will not be explicitly considered in this work.

CBRs often tend to conform into non-globular structures making it difficult to solve with usual experimental procedures [3], [4], [7]. Interestingly, several research groups have identified CBRs to be correlated (although not perfectly) to structural disorder [9], [10]. Regarding the functions of CBR proteins, there have been some early reports [3], [4], where a few examples of CBR proteins and their functions are presented. More recent works have focused on functional features of proteins with either specific types of homopolymeric runs [11] or (approximate) repeats [12] which, however, are not necessarily CBRs.

Two of the most popular approaches for identifying CBRs in amino acid sequences are SEG [4] and CAST [5]. SEG is essentially a two-pass algorithm based on information content (quantified using Shannon Entropy [13]): during the first pass, SEG identifies candidate CBRs with information content below a predefined threshold by scanning fixed-sized window segments of the query sequence; then, it optimizes the detected segments by merging neighboring candidates using a probabilistic approach and a more relaxed information content threshold. On the contrary, CAST is conceptually relying on the detection of unexpectedly high similarities of the query sequences to any of the possible homopolymeric amino acid sequences, using an iterative dynamic programming-based procedure. Early comparisons of these two algorithms [5], [14] illustrated the superiority of CAST when applied as a filter prior to BLAST searches; this finding can be clearly attributed to CAST's selective detection (and masking) properties compared to the more "aggressive" nature of CBR detection with SEG.

Our research goal is to investigate different properties of CBR instances in proteins with experimentally determined structures, in order to set the stage for more elaborate analyses of the structural impact of CBRs. Along these lines, we carefully collected representative protein sequence datasets and subsequently applied SEG and CAST with different settings for detecting CBRs. Herein, we report on the statistical features of sequential properties of CBRs, such as the frequency of proteins with a CBR (CBR proteins), the CBR length distributions and the types of CBRs.

II. Data and Methods

A. Protein Sequence Data

In the context of a larger project related to the detailed study of the structural properties of CBRs, we compiled a dataset consisting of high resolution protein structures solved by X-Ray crystallography from the Protein Data Bank (PDB, [15]). In our analysis it is essential to only include protein sequences sharing sequence identity below a certain threshold; this approach practically ensures that we will end up with a unique (or a few) representative(s) of each protein family, thus removing redundancy. PISCES is a protein sequence culling

[1]Albeit the fact that these terms are not necessarily synonymous, they will be used interchangeably throughout this manuscript.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

server, which offers subsets of protein sequences selected from the entire PDB according to structure quality and a maximum level of acceptable pairwise sequence identity [16]. We downloaded a pre-compiled non-redundant dataset from the PISCES website (2210 polypeptide chains after quality control; access date: 2/9/2009) using the following criteria: 30% sequence identity, resolution ≤ 1.6Å and R-factor ≤ 0.25.

As a more generic representative of the protein sequence universe, we used a UniProtKB [17] subset based on the reviewed (Swiss-Prot) entries, reduced for sequence similarity at the 50% pairwise identity level (UniRef50). This was achieved through the UniProtKB web site (139448 polypeptide chains; access date: 22/10/2010). Further reduction of UniProt entries to the same level of sequence identity with PDB entries was not practical in terms of computational workload and was consider a minor issue.

We have questioned the suitability of the dataset collected in a similar work [18], since we deemed that the sequence collection and redundancy removal procedures followed therein were not appropriate. Based on the description of the procedure given by these authors, we initially downloaded a set of PDB entries sharing no more than 30% sequence identity utilizing the "Advanced Search" tool on the PDB website, with a limit on "Deposit Date" up to 2/1/2007 (date of publication of [18]). Sequence chains in this dataset were further separated according to the experimental method used: X-Ray crystallography or NMR spectroscopy. The respective FASTA sequence files were also downloaded from the PDB website using the "FASTA file Download" tool (27604 and 1776 protein sequences, respectively). The NMR dataset was not used in our work, since we only concentrate on structures solved by X-ray crystallography. In order to perform the CBR analysis we further reduced the X-ray dataset (as instructed in [18]) by removing all chains with length below 41 amino acids and all proteins with only alpha carbon coordinates. We observed that this sequence dataset further contained sequences corresponding to RNA/DNA subunits[2] and we removed them as well. After these initial data cleansing steps the filtered dataset (namely, XBAN) contained 20452 sequences, and should be practically identical to the dataset used in [18]. XBAN was further reduced using a local installation of the PISCES software using the parameters described previously, and the reduced dataset (namely, RXBAN) contained only 1368 protein sequences, showing that the original XBAN dataset had a large amount of redundancy.

B. Detection of CBRs and CBR proteins

Sequence datasets were provided as FASTA formatted input files for SEG and CAST. For a given CBR detection scheme, we define a "CBR protein" as a protein with at least one detected CBR under the respective scheme. For the analysis reported herein we label each CBR by a residue type. While for CAST this label is inherent to the detection procedure [5], for SEG we post-process the results and assign a residue type to the most frequent residue type in the CBR.

Both methods were employed with different parameters (Table I) to investigate the effect of the different settings used to detect CBRs on the CBR properties. Importantly, a complication in the analysis stems from the existence of Histidine tags (His-tags), often used for affinity purification of recombinant proteins. In fact, their abundance in the PDB dataset skews the statistics regarding CBRs (data not shown), since more detection schemes would identify a stretch of His residues as a CBR. Thus, special care was taken to ignore His-tag segments from subsequent computation, without neglecting genuine histidine rich CBRs. Moreover, several instances of residues of undetermined type exist in sequences derived from the PDB: these residues are usually denoted with the same character used to mark residues in CBRs (i.e. 'X') and we had to also correct for this factor.

C. Statistical analysis

Standard statistical tests were employed as appropriate, using custom Perl code interfaced to existing CPAN Perl modules. More specifically, for the nonparametric Wilcoxon Rank Sum Test we used the Statistics::Test::WilcoxonRankSum module. A practical problem emerged when analyzing the huge amount of observations in the UniRef50 dataset; for this purpose we developed a bootstrapping version of this test. In particular, we performed the standard Wilcoxon test using 10000 sub-datasets composed by 10000 randomly selected samples with replacement from this dataset. Randomization was achieved by independently selecting (using the Math::Random module) 10000 random integers in the range $[1, \ldots, |UniRef50|]$, where $|.|$ denotes the cardinality of a set. In this setting, the null hypothesis (i.e. the two samples are drawn from a single population) is rejected at significance level α when the fraction β of tests yielding a p-value $< \alpha$ satisfies $\beta < 1 - \alpha$. We can interpret the quantity $1 - \beta$ as a bootstrap p-value.

Contingency matrices are used to examine the relationship between the observed and expected frequencies of two categorical variables, in our case the frequencies of amino acid residue types in or out of CBRs. For this purpose, we use a χ^2 test of independence with 1 degree of freedom (available from the Statistics::ChisqIndep module).

Assume the counts of residues from each residue type detected in CBRs by two detection schemes correspond to random variables $X = \{x_i\}, Y = \{y_i\}$, $i \in \{A, C, \ldots, Y\}$. For identifying correlations between the different levels of masked residue types detected by different CBR detection schemes, we calculate the Pearson Correlation Coefficient (PCC) $PCC_{xy} = \frac{\sum x_i y_i - n\hat{x}\hat{y}}{(n-1)s_x s_y}$, where $\hat{x}, \hat{y}$ and s_x, s_y are the observed means and standard deviations of the variables x, y and n represents the sample size. We use PCC as an intuitive measure of correlation, since $-1 \leq PCC \leq 1$, with $PCC = 1$ (-1) when x, y are perfectly correlated (negatively correlated).

[2]Actually, this is just one of the shortcomings of the sequence redundancy reduction methods offered by the PDB (for more details on this procedure, see: http://www.rcsb.org/pdb/statistics/clusterStatistics.do)

TABLE I: Parameter sets for SEG and CAST.

Algorithm	Parameters	Notes
SEG	L=6, K1 = 0, K2 = 0	SEG6: detects homopolymers of length 6;
		L denotes the window size, K2 (1) the trigger complexity and K2 (2) the extension complexity [4]
	L=7, K1 = 0, K2 = 0	SEG7: detects homopolymers of length 7
	L=12, K1 = 0, K2 = 0	SEG12a: detects homopolymers of length 12
		Results for modes SEG6, 7 and 12a are not reported, since they detected no CBRs after correcting for His-tags
	L=12, K1 = 2.2, K2 = 2.5	SEG12: mainly detects short CBRs, was BLAST default
	L=25, K1 = 3.0, K2 = 3.3	SEG25: detects medium length CBRs
	L=45, K1 = 3.4, K2 = 3.75	SEG45: detects longer CBRs
CAST	cutoff:15, 20, 25, 30, 35, 40	substitution matrix: BLOSUM62; masking modes named using the convention: CAST[cutoff]

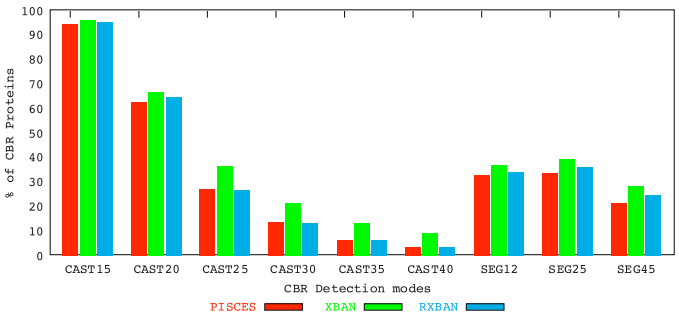


Fig. 1: Fraction of CBR proteins detected by different schemes. PISCES: the non-redundant dataset explicitly collected for this work; XBAN: the subset of X-ray structures from the recollected Bannen dataset [18]; RXBAN: our reduced version of the XBAN dataset.

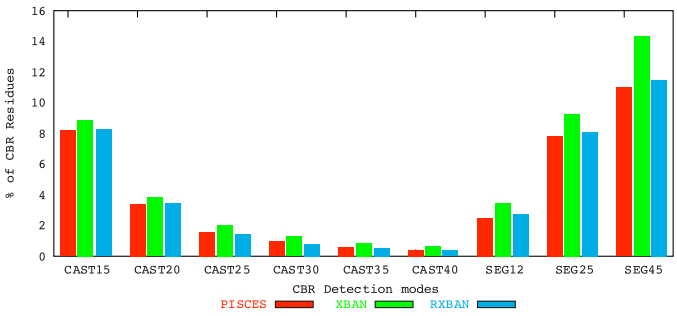


Fig. 2: Fraction of residues marked as CBR. See text and Fig. 1 for descriptions of the datasets.

III. Results

A. Frequency of CBR proteins

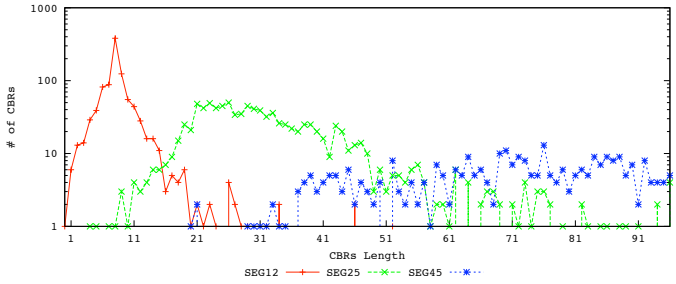
Inspecting CAST results, we observe an exponential decay of the fraction of CBR proteins as a function of detection cutoff; this observation holds both for the PISCES (this work) and the XBAN [18] datasets (Fig. 1). It is worth mentioning that the most permissive thresholds (i.e. 15, 20) detect most of the proteins as CBR. It is expected that, at least for CAST15 these results could reflect global compositional effects (Fig. 2 and section III-C). Respective figures are computed for SEG, with SEG25 detecting the highest fraction of CBR proteins compared to the other modes (Fig. 1). It also appears that CAST25 detects a similar fraction of CBR proteins when compared to all SEG modes, however, we have not investigated the overlap between the respective datasets. Notably, it is evident that all CBR detection modes consistently detect a higher fraction of CBR proteins in the XBAN dataset compared to PISCES (Fig. 1), which is probably related to the redundant nature of this dataset. When the same analysis was performed on the RXBAN dataset (redundancy reduced XBAN) the fraction of CBR proteins was practically identical to the PISCES dataset (Fig. 1). This strongly indicates that the XBAN dataset collected in [18] may not be the most appropriate for assessing the impact of low-complexity regions in protein structure determination.

When taking into account the absolute numbers of masked residues detected by the different schemes, it is evident that different CAST modes again follow an exponential decay pattern (Fig. 2). However, the figures are very different when SEG modes are taken into account. In particular, SEG45 (which detects less CBR proteins) filters a large fraction of residues (approx. 11% in the PISCES dataset). Clearly, this is due to the fact that this mode detects long CBRs. Importantly, regarding CAST15, which masks approx. 95% of the sequences, it only filters a negligible fraction of residues (approx. 8%), illustrating that even at this very permissive threshold, CAST masking largely remains highly selective. The respective figures for the RXBAN dataset were again practically identical to those for the PISCES dataset; therefore, all the following analyses focus into the PISCES dataset.

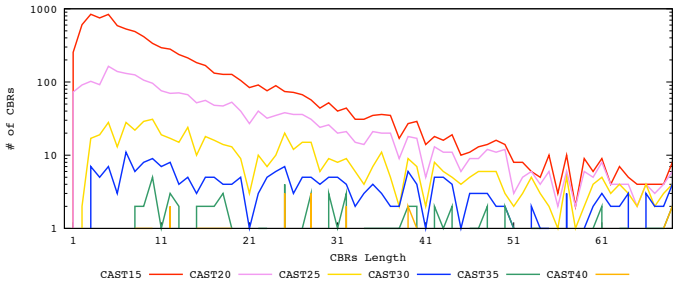
B. CBR length distributions in the PISCES dataset

The detailed CBR length distributions reveal that the window-based nature of SEG drives SEG-based modes to perform in a window length-dependent manner (Fig. 3a). On the other hand, we observe more similarities between the distributions for different CAST modes (Fig. 3b). The average and median lengths of CBRs detected in the PISCES dataset are shown in Table II.

Binning CBR lengths reveals that most of the CBRs detected by CAST modes are less than 80 residues long (Fig. 4a). It is worth mentioning that the maximum value of CBR length recorded in this dataset is 457 residues long, and was reported by all modes except CAST40. This extremely long CBR is a T-rich domain (CAST score: 38) spanning throughout the whole sequence (residues 5-461) of Internalin A (chain A, PDB ID: 1O6V), a protein with internal structural repeats. Another interesting, yet unexplained feature, is the sharp increase in the bin 91-100, which is observed for all cutoffs tested. This irregularity seems to mainly involve the most frequent CBR-



(a) SEG.



(b) CAST.

Fig. 3: Detailed CBR length distributions. Distributions have been truncated to display CBRs shorter than 96 (SEG) and 70 (CAST) residues for clarity. x-axis: CBR length; y-axis: number of CBRs in log-scale.

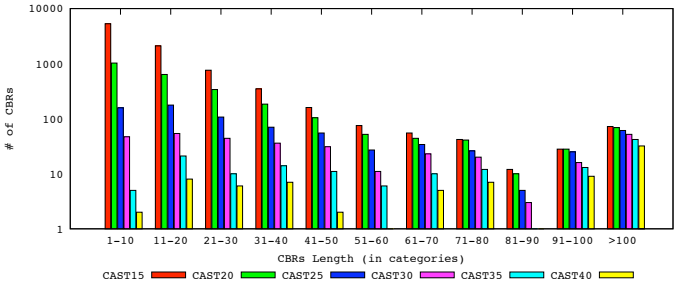
TABLE II: Descriptive statistics for CBR lengths.

Detection mode	No.CBR proteins	Average Length	Median Length	Standard Error (Mean)
CAST15	2090	14.12	9	0.21
CAST20	1386	23.52	13.5	0.65
CAST25	601	40.89	25	1.86
CAST30	300	58.77	37	3.62
CAST35	139	89.54	67	7.10
CAST40	77	109.95	92	9.96
SEG12	721	12.68	12	0.13
SEG25	747	37.80	34	0.51
SEG45	472	103.35	90.5	2.34

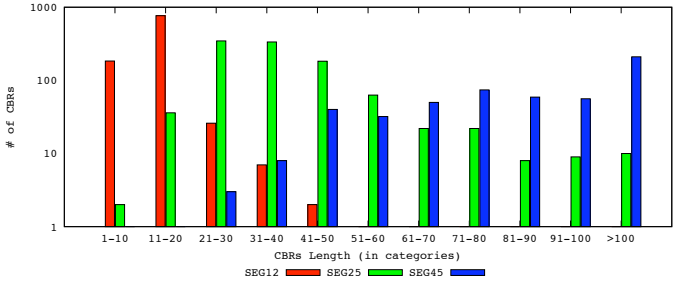
types (Ala-, Ser- and Thr-rich domains; data not shown). Nevertheless, more detailed inspection is necessary to resolve whether any biological significance exists in this finding.

An exponential decay pattern is observed for the number of CBRs as a function of CBR length for more relaxed CAST modes (CAST15-30). On the other hand, in SEG results it is evident that artifacts are generated due to its window-based detection scheme. More specifically, a clear peak for CBRs of length 11-20 is observed for SEG12 (window length=12). For the remaining modes this effect is still present, although to a lesser extent, probably due to the double-pass nature of SEG where overlapping candidate LCRs are merged.

In the detailed (un-binned) results (Fig. 3b), we observe a sharp peak at lengths 2-6 for the more permissive CAST thresholds (15 and 20). This observation could be initially attributed to runs of rare residues (e.g. Cys, Trp and His) with high self-matching scores in the BLOSUM62 matrix (C-C: 9,



(a) CAST.



(b) SEG.

Fig. 4: Binned CBR lengths. x-axis: CBR length; y-axis: number of CBRs in log-scale.

W-W: 11, H-H: 8). Detailed inspection of the results revealed that this was the case: for the short CBRs (≤ 4 residues) the number of CBRs detected for each type seems to be positively correlated to the self-score (Table III).

TABLE III: Short CBRs (≤ 4 residues long) for different residue types (CAST15). Bold and red characters denote rare residue types and residue types with high counts, respectively.

Residue (x)	Counts	BLOSUM62 x-x scores	Residue (x)	Counts	BLOSUM62 x-x scores
A	7	4	M	4	5
C	**211**	**9**	N	24	6
D	117	6	P	79	7
E	52	5	Q	14	5
F	89	6	R	42	5
G	200	6	S	0	5
H	**310**	**8**	T	32	5
I	10	4	V	19	4
K	21	5	**W**	**278**	**11**
L	12	4	Y	180	7

Regarding the distribution of CBR lengths for the UniRef50 dataset, we performed the Wilcoxon Rank Sum test (see section II-C for details), and significant differences were detected for all CBR detection modes at the 99.95% confidence level for the most relaxed CAST modes (15-30) and all tested SEG modes (bootstrap p-value $= 0$). Whereas for CAST35/40 the null hypothesis cannot be rejected (bootstrap p-value $= 1$), a finding which may propose that using these CBR detection settings the length properties of CBRs in the structural dataset cannot be distinguished from those of UniRef50. Therefore, assuming that CBR lengths in PDB should follow a different length distribution compared to the overall sequence database,

it may be appropriate to avoid using these CBR detection modes for studying CBRs in the PDB.

Further qualitative analysis was performed for the PISCES dataset, where we observe that certain CBR types (e.g. Ala, Ser, Thr, Gln, Lys) are detected throughout the length range, while others (e.g. Ile, Val, Tyr, Phe) are restricted to shorter CBRs only ($\leq$ 50 residues long). However, more elaborate analysis is necessary (and still under way) in order to link this interesting finding with possible structural properties of different CBR types.

C. Types of Compositionally Biased Regions

Analysis regarding the individual types of compositionally biased residues was performed for all CBR detection modes for the PISCES and UniRef50 datasets, in order to identify whether

- CBR results reflect global compositional properties of the datasets under study, and
- particular CBR types are favored/disfavored in experimentally determined 3D structures

We compared the background frequencies of the different amino acid residue types for the two databases: Ala, Glu, Gly, Leu, Ser and Val are the most abundant residues in both datasets and the rare residue types Cys and Trp are found with relative frequencies of approx. 1%. A χ^2 test of independence (99.9% confidence level with df = 19 degrees of freedom: $\chi^2 = 0.7347$, p-value = 1) indicates that the two distributions are not independent. In fact, there is a strong linear relationship between the relative frequencies in the two datasets ($y = 0.9757x + 0.0012$, $R^2 = 0.9083$), therefore, we can practically consider that the background distributions between the two datasets are comparable.

Using CAST, it is worth mentioning that certain CBR types are not being detected at all or are detected in very low numbers, this observation being more pronounced in less permissive thresholds (25-40). Such residues include mostly hydrophobic residues (Val, Ile, Leu, Met, Tyr, Phe), rare residues (Cys, Trp) and Arg. An interesting observation relates to Val and Leu, which despite the fact that they are among the most common amino acids, they are rarely found in CBRs. Even though we removed all His-tags (see section II-B), His is still relatively frequently detected in CBRs in the PISCES dataset. With all SEG modes we observe Ala, Gly, Leu and Val to be frequently labeled as CBR residues, closely resembling the global composition of the database.

In order to clarify to which extent average database composition relates to the types of residues detected in CBRs, it is necessary to have some quantitative estimates. Towards this end, using the average global composition of proteins in our datasets, we investigated whether it can be identified as a source for the detected CBRs. In particular, we performed a χ^2 test of independence (99.5% confidence level with df = 19 degrees of freedom) between the fraction of masked residue types and the global composition. With the exception of CAST15 (p-value=0.33) all other CAST modes demonstrated independence to the global composition (p-values $\leq$ 0.01). On the contrary, for all SEG modes the null hypothesis (dependence) cannot be rejected at this significance level (p-values $\geq$ 0.89). Therefore, it is evident that the composition of CBRs detected by SEG are expected to reflect the average residue content of the database.

A χ^2 test of independence (99.5% confidence level with df = 19 degrees of freedom) was performed for each pair of CBR detection modes. For tests involving CAST15 against all SEG modes the null hypothesis could not be rejected (data not shown); this mode demonstrated independence only against CAST35 and 40 (p-values: 2.95×10^{-2} and 1.73×10^{-4} respectively). Even though SEG12 and SEG25 behave quite similar to CAST25 in terms of the total number of CBR proteins, when the CBR types are examined CAST25 results are significantly different. In particular, for tests involving CAST25 against all SEG modes the null hypothesis is always rejected (p-values $\leq 4 \times 10^{-3}$). Similar results hold for all other CAST-vs-SEG tests, whereas for SEG-vs-SEG tests data do not support independence.

Using the Pearson Correlation Coefficient, we test whether any correlation exists between the results obtained by different masking modes. As a general trend, we observe that there is higher correlation between results obtained by the same method. For example, CAST25 has $PCC \geq 0.9$ when compared to the other CAST modes, whereas $0.57 \leq PCC \leq 0.73$ when compared to SEG modes.

IV. Conclusion

Global compositional characteristics of nucleic and amino acid sequences have been used as features to predict properties at multiple levels, from the molecular (e.g. function, expression, origin) to the organismic level (e.g. niche) [19]–[23].

However, only a few systematic efforts have been made so far in order to elucidate whether local compositional bias contains useful information which may be exploited for similar tasks [11], [24]. The main reasons for this shortage are:

- CBRs have been mostly dealt with as an undesirable sequence feature (e.g. masking prior to database searches, selecting against when seeking targets for structural genomics) and not as an informational character *per se*,
- divergent methods have been developed to detect local compositional bias, based on distinct definitions with largely different (and incomparable) results, and
- most of the aforementioned methods do not naturally classify CBRs into different types, thus prohibiting more detailed and informative analyses.

This work was an attempt to widen our understanding on the local compositional properties of amino acid sequences, with the future prospect of understanding their relation to the protein three dimensional structure.

The substantial differences in the performance of distinct CBR detection modes observed in our analysis opens new questions and avenues for research, with possible applications in target selection for current large-scale structural genomics projects. Despite the fact that we have illustrated that the dataset used in the a recent report [18] on the impact of

CBRs in protein structure determination suffers from extensive sequence redundancy, further work is necessary to illuminate whether the main findings of this analysis should be revised.

Our results indicate that SEG, the most commonly used method for detecting CBRs, may not be the most appropriate tool for attempting to decipher CBR to 3D structure relationships. In particular, it seems that with this method CBR length distributions are largely dominated by the effect of the window-size parameter. In addition, the CBR residue type distribution resembles the global database distribution, making it questionable whether it will be possible to effectively (i) identify local compositional features related to protein crystalizability or (ii) highlight subtle structural preferences of different CBR types using SEG.

Based on the correlation patterns of the residue types marked as CBR, we propose that CAST, and in particular CAST25, may be the most appropriate CBR detection mode (upon those tested in this study) for seeking structural properties of CBRs and CBR proteins. CBR length distributions seem to follow a consistently similar pattern, independently of the detection threshold employed. Moreover, the distribution of residues marked as CBRs clearly deviates from the background frequencies, thus capturing truly local compositional extremes.

Interestingly, using CAST we identified an unexpected (and still unexplained) depletion of hydrophobic residues in CBRs in the structural dataset, and in particular of Val and Leu which are among the most frequent residue types. There are several possible explanations, such as the relatively low incidence of transmembrane proteins in the structural database or the trend in the avoidance of large hydrophobic surface patches, which usually result in aggregation prone interfaces [25]–[27]. A starting hypothesis (currently under investigation within our group) is that CBRs of hydrophobic type could be associated with transmembrane domains or buried hydrophobic cores in globular proteins. Of similar interest is the detailed characterization of the structural environments of CBRs based on residue type and length, which may provide novel tools for predicting protein structural features from amino acid sequences.

ACKNOWLEDGMENT

We thank Prof. Konstantinos Fokianos (University of Cyprus) for discussions on the bootstrap variant of the Wilcoxon Rank Sum test, Dr Miguel Andrade (Max Delbrück Center for Molecular Medicine) for insightful comments, and the University of Cyprus (fund 3/311) for financial support.

REFERENCES

[1] S. Karlin, P. Bucher, V. Brendel, and S. F. Altschul, "Statistical methods and insights for protein and DNA sequences," *Annu Rev Biophys Biophys Chem*, vol. 20, pp. 175–203, 1991.

[2] S. H. White, "Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure," *Annu Rev Biophys Biomol Struct*, vol. 23, pp. 407–39, 1994.

[3] J. C. Wootton, "Sequences with 'unusual' amino acid compositions," *Current Opinion in Structural Biology,*, vol. 4, no. 3, pp. 413–421, 1994.

[4] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers & Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[5] V. J. Promponas, A. J. Enright, S. Tsoka, D. P. Kreil, C. Leroy, S. Hamodrakas, C. Sander, and C. A. Ouzounis, "CAST: an iterative algorithm for the complexity analysis of sequence tracts." *Bioinformatics*, vol. 16, no. 10, pp. 915–922, 2000.

[6] M. M. Alba, R. A. Laskowski, and J. M. Hancock, "Detecting cryptically simple protein sequences using the simple algorithm," *Bioinformatics*, vol. 18, no. 5, pp. 672–678, 2002.

[7] M. A. Huntley and G. B. Golding, "Simple sequences are rare in the Protein Data Bank," *Proteins*, vol. 48, no. 1, pp. 134–140, 2002.

[8] K. L. Sim and T. P. Creamer, "Protein simple sequence conservation," *Proteins: Structure, Function and Bioinformatics*, vol. 54, pp. 629–638, 2004.

[9] P. Romero, Z. Obradovic, and A. K. Dunker, "Folding minimal sequences: the lower bound for sequence complexity of globular proteins," *FEBS letters*, vol. 462, no. 3, pp. 363–367, 1999.

[10] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, vol. 11, no. 11, pp. 1453–9, Nov 2003.

[11] W. Haerty and G. B. Golding, "Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences," *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, vol. 53, no. 10, pp. 753–762, Oct 2010.

[12] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting, "Protein repeats: Structures, functions, and evolution," *Journal of Structural Biology*, vol. 134, no. 2-3, pp. 117 – 131, 2001.

[13] C. E. Shannon, "A mathematical theory of communication," *Bell Labs Technical Journal*, vol. 27, pp. 379–423, 1948, iD: 34.

[14] D. P. Kreil and C. A. Ouzounis, "Comparison of sequence masking algorithms and the detection of biased protein sequence regions," *Bioinformatics*, vol. 19, no. 13, pp. 1672–1681, 2003.

[15] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: redesigned web site and web services," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D392–401, Jan 2011.

[16] G. Wang and R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[17] UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D71–5, 2012.

[18] R. M. Bannen, C. A. Bingman, and G. N. Phillips Jr, "Effect of low–complexity regions on protein structure determination," *J Struct Funct Genomics*, vol. 8, pp. 217–226, 2007.

[19] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé, "Codon catalog usage and the genome hypothesis," *Nucleic Acids Res*, vol. 8, no. 1, pp. r49–r62, Jan 1980.

[20] A. Campbell, J. Mrázek, and S. Karlin, "Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA," *Proc Natl Acad Sci U S A*, vol. 96, no. 16, pp. 9184–9, Aug 1999.

[21] S. Karlin, J. Mrázek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *J Bacteriol*, vol. 179, no. 12, pp. 3899–913, Jun 1997.

[22] D. P. Kreil and C. A. Ouzounis, "Identification of thermophilic species by the amino acid compositions deduced from their genomes," *Nucleic Acids Res*, vol. 29, no. 7, pp. 1608–15, Apr 2001.

[23] D. Cortez, L. Delaye, A. Lazcano, and A. Becerra, "Composition-based methods to identify horizontal gene transfer," *Methods Mol Biol*, vol. 532, pp. 215–25, 2009.

[24] V. J. Promponas, "A simple clustering approach for pathogenic strain identification based on local and global amino acid compositional signatures from genomic sequences: the Escherichia genus case," in *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, nov. 2009, pp. 1 –4.

[25] S. Y. Patro and T. M. Przybycien, "Simulations of reversible protein aggregate and crystal structure," *Biophys J*, vol. 70, no. 6, pp. 2888–902, Jun 1996.

[26] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson, "Rationalization of the effects of mutations on peptide and protein aggregation rates," *Nature*, vol. 424, no. 6950, pp. 805–8, Aug 2003.

[27] R. Jacak, A. Leaver-Fay, and B. Kuhlman, "Computational protein design with explicit consideration of surface hydrophobic patches," *Proteins*, vol. 80, no. 3, pp. 825–38, Mar 2012.