BMC Bioinformatics

RESEARCH ARTICLE Open Access

Mining protein loops using a structural alphabet and statistical exceptionality

Leslie Regad[1*], Juliette Martin[2,3], Gregory Nuel[4], Anne-Claude Camproux[1]

Abstract

Background: Protein loops encompass 50% of protein residues in available three-dimensional structures. These regions are often involved in protein functions, e.g. binding site, catalytic pocket... However, the description of protein loops with conventional tools is an uneasy task. Regular secondary structures, helices and strands, have been widely studied whereas loops, because they are highly variable in terms of sequence and structure, are difficult to analyze. Due to data sparsity, long loops have rarely been systematically studied.

Results: We developed a simple and accurate method that allows the description and analysis of the structures of short and long loops using structural motifs without restriction on loop length. This method is based on the structural alphabet HMM-SA. HMM-SA allows the simplification of a three-dimensional protein structure into a one-dimensional string of states, where each state is a four-residue prototype fragment, called structural letter. The difficult task of the structural grouping of huge data sets is thus easily accomplished by handling structural letter strings as in conventional protein sequence analysis. We systematically extracted all seven-residue fragments in a bank of 93000 protein loops and grouped them according to the structural-letter sequence, named structural word. This approach permits a systematic analysis of loops of all sizes since we consider the structural motifs of seven residues rather than complete loops. We focused the analysis on highly recurrent words of loops (observed more than 30 times). Our study reveals that 73% of loop-lengths are covered by only 3310 highly recurrent structural words out of 28274 observed words). These structural words have low structural variability (mean RMSd of 0.85 Å). As expected, half of these motifs display a flanking-region preference but interestingly, two thirds are shared by short (less than 12 residues) and long loops. Moreover, half of recurrent motifs exhibit a significant level of amino-acid conservation with at least four significant positions and 87% of long loops contain at least one such word. We complement our analysis with the detection of statistically over-represented patterns of structural letters as in conventional DNA sequence analysis. About 30% (930) of structural words are over-represented, and cover about 40% of loop lengths. Interestingly, these words exhibit lower structural variability and higher sequential specificity, suggesting structural or functional constraints.

Conclusions: We developed a method to systematically decompose and study protein loops using recurrent structural motifs. This method is based on the structural alphabet HMM-SA and not on structural alignment and geometrical parameters. We extracted meaningful structural motifs that are found in both short and long loops. To our knowledge, it is the first time that pattern mining helps to increase the signal-to-noise ratio in protein loops. This finding helps to better describe protein loops and might permit to decrease the complexity of long-loop analysis. Detailed results are available at http://www.mti.univ-paris-diderot.fr/publication/supplementary/2009/ACCLoop/.

* Correspondence: leslie.regad@univ-paris-diderot.fr
[1]MTi, Inserm UMR-S 973, Université Paris Diderot- Paris 7, Paris, F-75205 Cedex 13, France

Background

Protein structures are classically described using secondary structures: α-helices, β-strands and loops, also called coils. This third class is a default description, which denotes all residues that are not involved in periodic local structures, helices or strands. On average, protein loops encompass 50% of residues. Protein loops are often involved in protein functions [1]. They participate in active sites of enzymes [2] and in molecular recognition [3,4]. They are often the place of binding sites: for example, the ATP and GTP-binding site (P-loop motif) and the calcium-binding site (EF-hand motif) are found in loops [5-8]. The description and analysis of protein loops have been the subject of many studies. Protein loops were first seen as random because they are highly variable in terms of sequence and structure and are subject to frequent insertions and deletions [9,10]. Because of their large variability, loops are the protein regions which are the most difficult to analyze and modelize. Indeed, in protein models, loops, and more particularly long loops, are the place of a lot of errors.

Systematic studies actually showed that loops, even long ones, are far from random. In their study, Panchenko et al. (2004) analyzed the evolution of protein loops and identified a linear correlation between sequence similarity and average loop structural similarity in protein families [11]. They suggested that the evolution of loops is made *via* an insertion/deletion process and concluded that even longer loop regions cannot be defined as "irregular conformations" or "random coils".

The resolution of an increasing number of protein structures allowed the classification of short loops (3 to 12 residues) according to their geometry, and gave birth to several loop classification systems: `Sloop` [12-14], `Wloop` [15,16], `ArchDB`[17-19], Li et al. classification [20,21]. These different classification initiatives were based on different criteria such as loop length [12,14,15,17,18], flanking region type [12,14,17,18,20,21], flanking-region geometry [12,14,17,18], or loop conformation [17,18]. The majority of the resulting loop clusters presented a significant sequence signature. These classifications thus revealed the existence of recurrent loop conformations with amino-acid dependence. However, these classifications focus on short and medium loops (less than 12 residues) and do not take long loops into consideration.

Another type of studies focused on specific structural motifs extracted from loops such as β-turn [22-25], β-hairpin [26-29], helix-turn-helix [30], helix-turn-strand [31], or ω-loop [1,32]. The most frequent motif is β-turn. It corresponds to 25% of residues [33]. Other turn types have been identified such as γ-turn [34-36] or α-turns [37,38]. Recently, Golovin et al. (2008) proposed a web application that allows identifying known small structural motifs characterized by hydrogen-bonds (alpha-beta motif, asx-motif, beta-bulge, beta-bulge-loop, beta-turn, catmat, gamma-turn, nest, schellmann-loop, st-motif, st-staple, st-turn) from a query protein [8]. A database of these structural motifs extracted from a set of 400 representative proteins is now available [39]. All these studies were dedicated to particular -and known-small structural motifs, but did not perform a systematic analysis of all loops.

In a previous study, we have shown that the structural alphabet HMM-SA (Hidden Markov Model-Structural Alphabet) is an effective tool to simplify loop structures with good accuracy [40]. Structural alphabets constitute a privileged tool to discretize 3D structures including loop regions, with an accuracy that depends on the size of the fragment library [41]. HMM-SA is a collection of 27 structural prototypes of four residues called structural letters, permitting the simplification of all three-dimensional (3D) protein structures into uni-dimensional (1D) sequences of structural letters [42].

Here, we present an extensive analysis and description of both short and long loops based on the analysis of structural motifs extracted from loops. The systematic extraction of seven-residue structural motifs is based on the loop decomposition in structural letters provided by HMM-SA. Thanks to this decomposition, structural motifs are described as patterns of structural letters, called structural words. This representation as structural words permits to partition the full space of loop conformations, independently of their length, in clusters represented by distinct words. We first present general results concerning structural words: repartition of clusters and intrinsic characteristics of structural words such as structural variability and sequential specificity. Then, we present the analysis of the link between structural words and loop types. In order to gain further insight into the high complexity of loop structures, we complement our analysis with an original approach based on statistical exceptionality implemented in the SPatt software [43]. The idea is to compute, for each structural motif, a score that is a measure of its "unusualness" with respect to some background model. The goal is to assess whether some structural motifs are more or less frequent than expected. This is directly inspired by analogous studies of sequence patterns in genomes [44,45], that permitted the discovery of functional patterns such as restriction sites [46], cross-over hot spot instigator sites [47] and polyadenylation signals [48]. Finally, this systematic structural-alphabet decomposition and word analysis provide an accurate description of loops and allows extracting meaningful motifs in both short and long loops, which is an important contribution to the difficult task of long loop analysis.

Results

We extracted all structural motifs within loops from a non-redundant data set of 8186 protein chains, using the structural alphabet HMM-SA. This alphabet is a collection of 27 prototypes of four residues, denoted [A-Z, a], based on a hidden Markov model [40,42]. It permits the encoding of a protein structure of n residues into a sequence of (n - 3) structural letters.

Loop structures extracted from our protein data set were encoded into structural-letter sequence using HMM-SA. Each encoded loop was then decomposed into overlapping structural words, i.e. series of k consecutive structural letters, corresponding to k - 3 residue fragments. Thus, structural words can be seen as a way of clustering the fragments. Each cluster of fragments is defined by a structural word. The first step of this work is the determination of the optimal length of fragments/words.

Choice of the structural word length

The choice of the optimal length was guided by the following dilemma. On the one hand, it is desirable to consider long fragments, in order to better describe 3D conformation and capture the longest-range interactions. On the other hand, the amount of available data rapidly becomes insufficient when dealing with long fragments. To choose this optimal length, we computed the frequency of all structural words in our data set, with length from five residues (two-structural letters) to ten residues (seven-structural letters), see Additional file 1. We identified seven residues as the maximum length to avoid the problem of data sparsity. The number of different structural words sharply increases beyond that limit and 80% of structural words of 8 residues are seen at most 6 times in our data set, versus 34 times for words of 7 residues. For these reasons, we selected seven residues, i.e., four structural letters as the most meaningful length for systematic extraction.

First Part: Global results on structural words

We systematically extracted structural words of four structural-letters from protein loops and analyzed their properties: structural variability, amino-acid specificity and preference for particular loop types.

Extraction of structural words from loops

The data set contained 93396 loops of minimal length seven residues (i.e. four structural letters). From these loops, we extracted 415071 overlapping seven-residue fragments. The 415071 fragments were partitioned into 28274 different four-structural-letter words, with an average cluster size of 14.7 and a high variability: standard deviation was equal to 36. As HMM-SA offers a very detailed description of loop structures, some slightly different conformations ended up in distinct clusters; our classification then disclosed with a high number (5626) of singletons, i.e. clusters containing only one fragment. However, even if we had considered X-ray structures with good resolution (better than 2.5 Å), such rare conformations might have been an artifact due to the structural flexibility of some protein regions. Indeed, protein loops are generally more flexible than regular secondary structures [49]. We tested this hypothesis using B-factors, as atoms with high B-factors are those with the largest positional uncertainty. We computed the average Cα B-factor for all fragments in each structural word. We used the rule-of-thumb suggested in [50] and set a B-factor cut-off at 40. We found that a large proportion (28%) of singletons have an average B-factor greater than 40, compared to only 1% for structural words from clusters with more than 30 fragments. Singletons and rare conformations are thus linked to structural flexibility. In the rest of the paper, we consider a restricted set containing words seen more than 30 times (i.e., minimal cluster size set to 30), denoted $W\text{set}_{\geq 30}$. The reason for this choice is that our goal is to perform a statistical analysis of word properties, namely structural variability and sequence specificity. Since these properties are assessed by RMSd and Z-scores extracted from sequence profiles, a sufficient number of fragments per cluster is needed. We estimated that 30 fragments were sufficient to compute mean RMSd and sequence profiles. Statistics of $W\text{set}_{\geq 30}$ are given in Table 1. As can be seen in Table 1, $W\text{set}_{\geq 30}$

Table 1 Quantification of the structural word extraction from the non-redundant data set.

Words	$W\text{set}^{1}_{\geq 30}$	UR^{2}_{w}	NS^{3}_{w}	OR^{4}_{w}
Number of words	3310	166	2214	930
(%)	(11.7%)	(5.0%)	(66.9%)	(28.1%)
Number of fragments	249953	11435	129781	108737
(%)	(60.2%)	(4.6%)	(51.9%)	(43.5%)
Nb fragments/word	75.5	68.9	58.6	116.9*
All-loop coverage rate	72.7%	5.1%	46.5%	40.2%
Short-loop coverage rate	70.3%	4.4%	38.9%	39.3%
Long-loop coverage rate	74.9%	5.7%	53.9%	41.1%
Loops containing at least one word	84.8%	9.8%	60.3%	58.2%
Short loops containing at least one word	79.7%	6.1%	48.1%	49.4%
Long loops containing at least one word	97.8%	19.1%	90.9%	80.4%

1: words seen more than 30 times. 2: under-represented words, 3: non-significant words, 4: over-represented words, '*': significantly higher occurrence according to a Kruskal-Wallis test. Coverage rates are given on a per structural letter basis. Numbers within brackets denote the percentage of words/fragments with respect to the 28274 words/415071 fragments of the whole data set (column 1) and with respect to the 3310 words/249953 fragments in $W\text{set}_{\geq 30}$ (columns 2 to 4).

encompass 3310 different structural words (12% of all words), and 60% of fragments.

Loop coverage by Wset$_{\geq 30}$ words

In this part, we check if the elimination of rare words does not result in (i) a dramatic diminution of loop coverage or (ii) a loss of diversity in structural families.

At first, we can observe that the selection of *W*set$_{\geq 30}$ words does not favor any loop length: the distribution of loop lengths in *W*set$_{\geq 30}$ is similar to the global loop-length distribution (cf. Additional file 1).

Global loop coverage (cf. Materials and Methods). Words from *W*set$_{\geq 30}$ encompass 60% of the fragments. However, since we extracted overlapping fragments, the coverage rate of loop structures is more than 60%: if a loop of 8 structural letters is described by two *W*set$_{\geq 30}$ words on positions 1 to 4 and 5 to 8, the actual coverage is 100% even if only 2 out the 5 overlapping fragments are represented by frequent words.

Coverage rates are reported in Table 1. The limited number of words seen more than 30 times (3310) covers most loop, namely 73% of loop lengths. If we make the distinction between short loops (up to 12 residues) and long loops (longer than 12 residues), we can see that *W*set$_{\geq 30}$ words cover both short and long loops. If we now consider loops that contain at least one *W*set$_{\geq 30}$ word, *W*set$_{\geq 30}$ words partially describe 85% of all loops -80% of short loops and 98% of long loops.

The consideration of the restricted set *W*set$_{\geq 30}$ thus allowed us to get rid of clusters with high positional uncertainty while still covering a large fraction of protein loops.

SCOP superfamily coverage by Wset$_{\geq 30}$ words There might be a risk that the selection of recurrent words could give preferences to loops from highly populated structural families. In order to address this problem, we assessed the coverage of *W*set$_{\geq 30}$ with respect to the SCOP classification. We surveyed the SCOP classification of 8140 protein chains covered by *W*set$_{\geq 30}$. The results are presented in Table 2. We identified 1493 different superfamilies in the full data set. The removal of rare words led to the elimination of 46 protein chains, and 11 SCOP superfamilies. We then checked the number of structure members in the 1485 remaining superfamilies. After the removal of words seen less than 30 times, this number was lowered for 46 superfamilies. The majority of affected superfamilies (44 among 46) lost only one member, as shown in Additional file 1. These elements suggest that the elimination of words seen less than 30 times still permits to keep a good representation of SCOP superfamilies, since 97% of initial superfamilies were unaffected. Therefore, loops from highly populated structural families are not given preferences due to the selection of recurrent words.

Table 2 Population of SCOP superfamilies before and after elimination of rare conformations

	All words	*W*set$_{\geq 30}$
Nb structures	8186	8140
Nb superfamilies	1493	1485
Nb superfamilies with less than 30 members	1437	1435

Consequently, we can conclude that the systematic extraction of structural words shows that most loops can be described by a limited number of frequent four-structural-letter words.

Structural and amino-acid conservation of words

The next step consists in analyzing the intrinsic structural and sequential properties of structural *W*set$_{\geq 30}$ words. We considered the following properties: structural variability of the fragments, and dependence to their amino-acid sequence.

Structural properties of words The intra-word structural variability of clusters is assessed using the average Root Mean Square deviation (RMSd$_w$) between fragments within the same cluster. The global mean RMSd$_w$ is equal to 0.85 Å (cf. Table 3). Words exhibiting the largest structural variability include structural letters J or F. It was expected because these two letters are the most structurally variable ones [42]. We can observe that the word structural variability could be quantify by the structural-letter type. This allows avoiding the computation of RMSd and the superimposition of word fragments. This analysis shows that most words exhibit a weak structural variability.

Table 3 Structural and sequential properties of words in *W*set$_{\geq 30}$ according to the statistical word type

Words characteristic	*W*set$_{\geq 30}$	*UR*$_w$	*NS*$_w$	*OR*$_w$
Average RMSd$_w$ (Å)	0.85	0.94	0.89	0.74*
(± standard deviation)	(± 0.4)	(± 0.4)	(± 0.4)	(± 0.3)
Average RMSd$_{dev}$ (Å)	2.72	2.67	2.69	2.76
(± standard deviation)	(± 0.6)	(± 0.6)	(± 0.6)	(± 0.7)
Average$Z_{\max}$	10.3	9.5	8.8	14.0*
(± standard deviation)	(± 6.1)	(± 3.8)	(± 4.0)	(± 8.4)
Average nb$_{pos*}$	3.3	3.0	2.9	4.1*
(± standard deviation)	(± 1.8)	(± 1.7)	(± 1.6)	(± 1.8)
Average $d^{Z\text{-score}}$	31.1	29.0	27.4	39.5*
(± standard deviation)	(± 9.7)	(± 5.7)	(± 5.3)	(± 13.8)

The upper part of the table corresponds to the analysis of word structural properties. The intra-word structural variability is analysed using the Root Mean Square deviation (RMSd) between fragments corresponding to the same word (RMSd$_w$). The inter-word structural variability is analysed using the RMSd between fragments of two different words (RMSd$_{dev}$). The lower part of the table corresponds to the analysis of sequential properties of words. The intra-word amino-acid preferences of a word are analysed using $Z_{\max}$ criterion (cf. Method section) and the number of significant position of a word (nb$_{pos*}$). The coverage of sequential space is analysed using the Euclidian distance between *Z*-score vectors (cf. Method section) ($d^{Z\text{-score}}$). Numbers within brackets indicate standard deviations. *: significant differences according to the Kruskal-Wallis test. The RMSd$_{dev}$ are computed on a subset of 890 words of *W*set$_{\geq 30}$. ^{a}: words shared by long and short loops.

Amino-acid preferences of words Intra-word amino-acid specificity is assessed using Z-score computation as described in Material and Methods. Briefly, we computed Z-scores for the 20 amino acids at the 7 positions of a structural word. We then considered the maximum Z-score, denoted $Z_{\max}$, measuring the strongest amino-acid specificity, and the number of significant positions, denoted $\mathrm{nb}_{\mathrm{pos}^*}$, indicating how many positions exhibit significant sequence specificity. As shown in Table 3, the global average $Z_{\max}$ (resp. $\mathrm{nb}_{\mathrm{pos}^*}$) is equal to 10.3 (resp. 3.3). Almost every word (97%) present at least one significant position ($Z_{\max} \geq 4$) and 19% of words have at least one very significant position ($Z_{\max} \geq 14$). Conversely, only 3% of words (89 words covering 2% of loops) have no informative position. Among the sequence-informative words, 198 words (6% of recurrent words) are highly informative, as all their positions are significant. These very informative words cover 16% of loops. Words with high $Z_{\max}$ contain structural letters `D` and `S`, in agreement with the fact that these two letters have very strong sequence specificity [51]. Thus we can conclude that most loops are composed of motifs with amino-acid specificities.

Correlation between structural variability and sequential specificity We can note that there is no obvious link between $Z_{\max}$ and RMSd_w (Pearson coefficient is equal to 0.09, cf. Additional file 1). The structurally less variable words are not systematically the most informative ones in terms of amino acids. Some words with high RMSd_w are informative in terms of sequence, as illustrated by word `FFFF`, with an RMSd_w equal to 2.5 Å and $Z_{\max}$ equal to 15.8 (an illustration of the word geometry is presented in Figure 1).

2590 words are characterized by both low structural variability and significant sequential specificity, with RMSd_w lower than 1 Å and $Z_{\max}$ greater than 4. These structural words cover 63% of loop regions. We can conclude that most loops are composed of motifs with a weak variability and amino-acid specificities.

Relation between structural words and loop type

After exploring the intrinsic structural and sequential properties of structural words, we analyzed their relationship with different loop types seen in proteins. We defined different loop-types according to their lengths and flanking secondary-structures [14,15,17,18].

Loop length We used the Kullback-Leibler asymmetric divergence, denoted KLD criterion [52] (cf. Methods) to extract the words that are significantly more frequent in long loops than expected. These words are classified as specific to long loops. Words specific to short loops are extracted in a similar manner. The result of this analysis is presented in Table 4. We found that 758 words (23% of $W\mathrm{set}_{\geq 30}$) are specific to long loops and 476 words (14% of $W\mathrm{set}_{\geq 30}$) are specific to short loops. It means that roughly one third of the structural words display a significant preference for a length range, and two thirds are unspecific, i.e., shared by short and long loops. In Table 4, we also reported the loop coverage achieved by words specific to short and long loops. It can be seen that half of loops are covered by words shared by long and short loops. About one third of short loops (resp. long loops) are covered by words specific to short (resp. long) loops.

Flanking regions We now consider the four possible flanking regions for a loop: $\beta\beta$: loops linking two β-strands, $\alpha\beta$: loops linking an α-helix and a β-strand, $\alpha\alpha$: loops linking two α-helices and $\beta\alpha$: loops linking a β-strand and an α-helix. We found that about 60% of $W\mathrm{set}_{\geq 30}$ display a significant preference for one of the four-flanking-region types. This word set permits to cover about 59% of loops. Thus, about half of the loops are described by flanking-region-specific words.

Loop length × flanking regions We then combine the loop length and loop type descriptors to distinguish eight types of loops. According to the KLD criterion, 2543 words (80% of $W\mathrm{set}_{\geq 30}$) exhibit a significant preference for one of the eight loop-types. This significant word set covers more than half of the loops (66%).

The association between words and the eight loop types is further explored using a correspondence analysis presented in Figure 2. The first two axes of the correspondence analysis capture 62% of the variability and are mainly explained by the preference for short loops. The $\beta\beta$ short loops is opposite to the $\alpha\alpha$ short loops on the first axis (36% of the variability) while the $\alpha\beta$ short loop is opposite to the $\beta\alpha$ short loop on the second axis (26% of variability). Association is weaker for long loops -appearing in the central region of the plot- but similar tendencies are observed for short and long loops. This analysis made it possible to identify the loop structures with a dependence to loop-type, and the ones with no dependence.

Loop-type preferences × intrinsic properties

By combining the loop-type preferences of words and their intrinsic properties, we observe that words specific to short loops present slightly higher sequence dependence than others, while words specific to long loops have lower structural variability (cf. Table 5).

We can note that only 44 words (1% of the $W\mathrm{set}_{\geq 30}$ words) have neither amino-acid-significant position, nor loop-type preference. Thus, less than 1% of loop regions are covered by these unspecific words in terms of sequence dependence and loop types.

Our approach, which relies on a systematic decomposition of short and long loops, allowed showing loops are composed of recurrent structural motifs, some of them with preference for a particular loop type in terms of loop length and/or flanking regions. Conversely, some

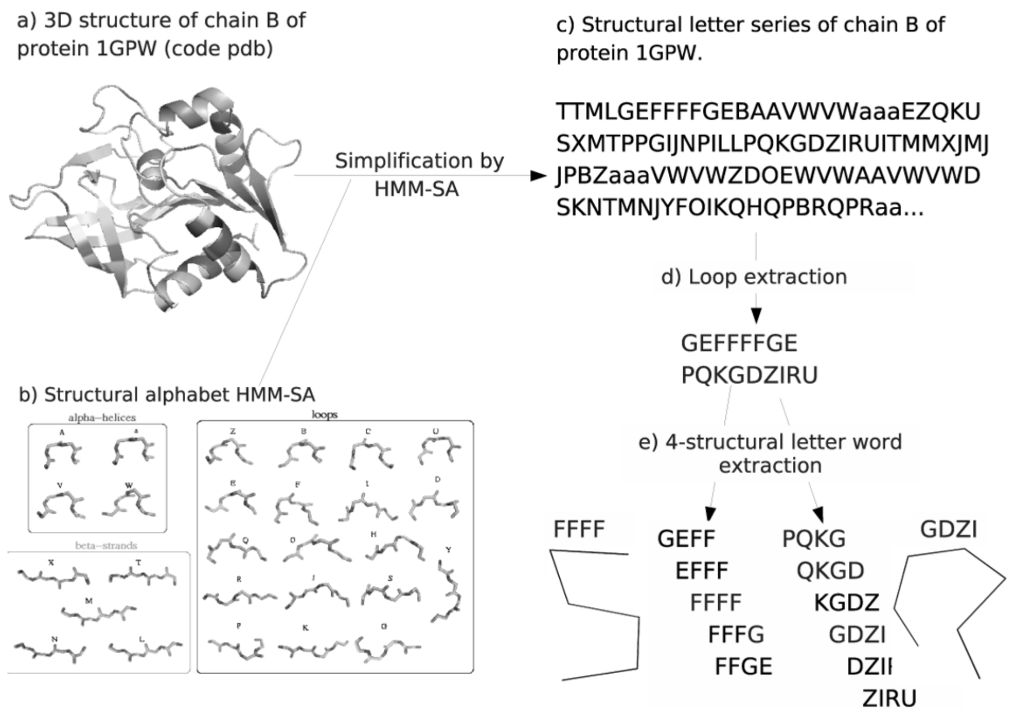
a) 3D structure of chain B of protein 1GPW (code pdb)
Simplification by HMM-SA
c) Structural letter series of chain B of protein 1GPW.
TTMLGEFFFFGEBAAVWVWaaaEZQKU
SXMTPPGIJNPILLPQKGDZIRUITMMXJMJ
JPBZaaaVWVWZDOEWVWAAVWVWD
SKNTMNJYFOIKQHQPBRQPRaa...
d) Loop extraction
GEFFFFGE
PQKGDZIRU
b) Structural alphabet HMM-SA
alpha-helices
loops
beta-strands
e) 4-structural letter word extraction
FFFF
GEFF
EFFF
FFFF
FFFG
FFGE
PQKG
QKGD
KGDZ
GDZI
DZII
ZIRU
GDZI

Figure 1 Loop-word extraction from chain B of protein 1GPW. a) 3D structure of the protein, b) the 27 structural letters of HMM-SA, c) structure simplification as a succession of structural letters, d) extraction of simplified loops, e) extraction of overlapping words of four structural-letters with structural words `FFFF` and `GDZI` illustrated.

structural words have no preference for a loop length, meaning that they are similarly found in short and long loops.

Second part: Statistical exceptionality of structural words

In the second part of this study, we complement our analysis of word properties by their statistical exceptionality in protein structures represented by strings of structural letters. Statistical exceptionality is traditionally used in genome analysis to extract functional motifs such as enzyme restriction sites or regulatory motifs [44-48]. Our goal was to explore if a statistical bias is also associated to specific properties in the case of protein structures. Statistical exceptionality does not measure the frequency of a word. It is an indicator of the discrepancy between observed and expected occurrence according to a background model that takes into account the first order Markovian process between structural letters. The statistical representation of words was assessed using the SPatt software that computes an exceptionality score L_p for each word (see Material and Methods). According to the value of L_p, words are classified as over-represented, under-represented or not significant. Hereafter, over-represented words are referred to OR_w, under-represented words as UR_w and not significant words as NS_w.

Table 4 Preference of structural words in *W*set$_{\geq 30}$ according to the loop types as assessed by the KLD criterion and the associated loop coverage rate (on a per structural letter basis).

Loop words specificity	*W*set$_{\geq 30}$[7]	UR$_w$[7]	NS$_w$[7]	OR$_w$[7]	all loops[4]	short loops[5]	long loops[6]
Long-loop-specific words	758 (22.9%)	23	475	260	23.2%	12.2%	33.9%
Short-loop-specific words	476 (14.4%)	23	220	233	25.7%	33.0 %	18.6%
Shared words 1	2076 (63.7%)	120	1519	437	45.9%	39.4%	56.3%
Flanking-region-specific words[2]	1879 (57.1%)	102	1131	646	58.6%	58.9%	58.4%
Flanking-region-unspecific words	1431 (43.2%)	64	1083	284	31.4%	21.7%	40.8%
Loop-type-specific words[3]	2543 (78.8%)	124	1605	814	66.3%	64.3%	68.2%
Loop-type-unspecific words	767 (23.2%)	42	609	116	16.6%	12.7%	20.5%

1: words shared by long and short loops 2: description by the four possible flanking-types ($\alpha\alpha$, $\alpha\beta$, $\beta\alpha$, $\beta\beta$ loops), 3: description by the four flanking-types and the two length ranges ($\alpha\alpha_s$, $\alpha\beta_s$, $\alpha\beta_s$, $\beta\beta_s$ for short loops and $\alpha\alpha_l$, $\alpha\beta_l$, $\alpha\beta_l$, $\beta\beta_l$ for long loops). 4: all-loop coverage rate (on a per structural letter basis) 5: short-loop coverage rate (on a per structural letter basis) 6: long-loop coverage rate (on a per structural letter basis) 7: Number of words

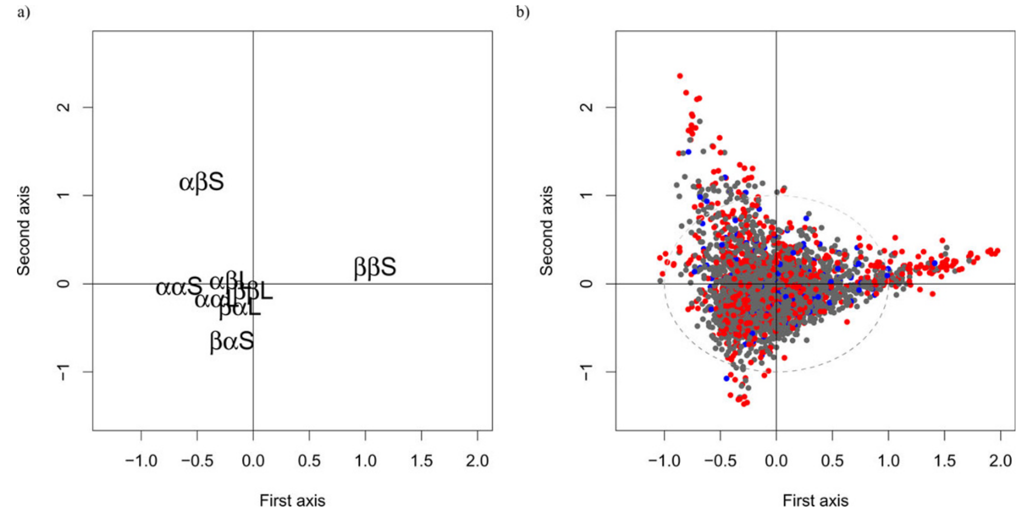

Figure 2 Correspondence analysis between the eight loop-types (defined according the length and the flanking regions of the loops) and the structural words in $W\text{set}_{\geq 30}$. $\alpha\alpha_s$, $\alpha\beta_s$, $\beta\alpha_s$, $\beta\beta_s$ correspond to the four different short-loop-types according to the flanking regions, and $\alpha\alpha_l$, $\alpha\beta_l$, $\beta\alpha_l$, $\beta\beta_l$ correspond to the four different long-loop-types according to the flanking regions. $\alpha\alpha$: loops linking two α-helices, $\alpha\beta$: loops linking an α-helix and a β-strand, $\beta\alpha$: loops linking a β-strand and an α-helix, $\beta\beta$: loops linking two β-strands. The two first axes account for 36% + 26% = 62% of the variance. a) Plot of the eight loop types, b) Plot of $W\text{set}_{\geq 30}$ words colored according to their statistical exceptionality: red = OR_w, gray = NS_w, blue = UR_w.

Extraction of exceptional words

The analysis of the correlation between the frequencies (i.e. cluster size) and L_p values for all words in the data set shows that many frequent words tend to be over-represented but there is no linear relation between frequency and exceptionality (cf. Additional file 1). Some frequent words are classified as UR_w or NS_w, like FFFF (seen 537 times, L_p = -2.8). Conversely, some rare words are classified as OR_w, like GDZI (seen 64 times, L_p = 102.2). An illustration of the geometry of these words is presented in Figure 1. This result shows the relevance of the extraction of word exceptionality instead of word frequency.

The repartition of words in $W\text{set}_{\geq 30}$ according to exceptionality status is given in Table 1. We can see that OR_w contribute predominantly to the set of fragments in $W\text{set}_{\geq 30}$: 40% of the fragments are in OR_w clusters. OR_w clusters are indeed significantly bigger than other word types (cf. Table 1).

Redundancy of loops and robustness of the extraction method

In this study, loops were extracted from a non-redundant data set presenting less than 50% sequence identity. Different redundancy levels have been used in the literature. Concerning loop classifications, Wloop[16] used a protein data bank with 50% sequence identity. The loop

Table 5 Structural and sequential properties of loop-type specific words in $W\text{set}_{\geq 30}$

Words characteristic	shared words[a]	short-long specific words	long-long specific words	flanking-region unspecific words	flanking-region specific words
Average RMSd_w (Å)	0.90	0.80*	0.74	0.83	0.88
(± standard deviation)	(± 0.4)	(± 0.4)	(± 0.4)	(± 0.4)	(± 0.4)
Average$Z_{\max}$	9.4	15*	9.7	8.5	11.6*
(± standard deviation)	(± 4.7)	(± 9.5)	(± 5.3)	(± 3.9)	(± 7.0)
Average $\text{nb}_{\text{pos}*}$	3.0	4.6*	3.3	2.7	3.8
(± standard deviation)	(± 1.6)	(± 1.7)	(± 1.8)	(± 1.5)	(± 1.8)

The upper part of the table corresponds to the analysis of word structural properties. The intra-word structural variability is analysed using the Root Mean Square deviation (RMSd) between fragments corresponding to the same word (RMSd_w). The inter-word structural variability is analysed using the RMSd between fragments of two different words (RMSd_{dev}). The lower part of the table corresponds to the analysis of sequential properties of words. The intra-word amino-acid preferences of a word are analysed using $Z_{\max}$ criterion (cf. Method section) and the number of significant position of a word ($\text{nb}_{\text{pos}*}$). Numbers within brackets indicate standard deviations. *: significant differences according to the Kruskal-Wallis test. The RMSd_{dev} are computed on a subset of 890 words of $W\text{set}_{\geq 30}$. [a]: words sharing by long and short loops classification system ArchDB is available in two versions: one built on a set of proteins with 40% sequence identity and the second on a redundant-protein set with 95% sequence identity [18]. It is classically considered that the evolutionary relationship between two proteins is detectable up to 25% sequence identity. Consequently this cut-off is frequently used for calibrating prediction methods [53]. Since loops are more variable than the rest of the protein sequence, we set the identity cut-off at 50% in order to work with as many data as possible with limited redundancy.

One could object that no attention was given to how many redundant loops were left or removed from the database during the redundancy filtering. The problem of loop redundancy is a non-trivial one: the extraction of loops from a non-redundant protein set does not necessarily result in a non-redundant loop set, and loop redundancy is itself difficult to quantify. We indirectly addressed this question by repeating our systematic extraction on different data sets, using identity levels of 25% and 80%. It was also important to ensure that our observations were applicable to protein structures in general and not only to the data set used. Taking into account the correction due to the different database sizes (see Method), we found a satisfactory level of consensus equal to 82% between the 25% and 50% databases, and 90% between the 50% and the 80% databases (more details are given in Additional file 1). These ratios refer to the proportion of recurrent words - common to both data sets - that are classified in the same statistical word type (over-presented/not significant/under-represented). Moreover, only one word, QLHB, was assigned as over-represented in a data set and under-represented in the other. Therefore, we can conclude that the extraction of exceptional words is robust and very weakly depends on the redundancy of the data set. Then, we compared the properties of the $W\text{set}_{\geq 30}$ words after classification into these three classes.

Exceptionality and word properties

The structural and amino-acid property measures for the three statistical word types (OR_w, NS_w and UR_w) are reported in Table 3.

The intra-word structural variability is lower for OR_w than for other words, as assessed by a Kruskal-Wallis test [54] (p-value < 2×10^{-16}, cf. Table 3). The RMSd_w distribution for the three statistical word types is shown in Figure 3a. It can be seen that the RMSd_w distribution of OR_w is shifted toward lower values. OR_w are thus significantly less structurally variable than other words.

The coverage of the structural space by the structural words of different exceptionality status is assessed by the RMSd between clusters. The goal is to evaluate how well the structural words sample the conformational space of loops. In order to assess the coverage of the loop-conformational space, we computed the RMSd between all pairs of words in the $W\text{set}_{\geq 30}$, denoted RMSd_{dev}. The average RMSd_{dev} computed for each type of words is given in Table 3. The average RMSd_{dev} for words in $W\text{set}_{\geq 30}$ is equal to 2.7 Å It is significantly greater than the average RMSd_w, indicating that the structural variability of words is low compared to the structural differences between words. This observation stands for the three types of words. RMSd_{dev} were computed between every words of $W\text{set}_{\geq 30}$, and the resulting 3310 × 3310 dissimilarity matrix is used to compute Sammon's map projections shown in Figure 3b. It can be seen that the three statistical word types all sample the conformational space in the same way. It means that OR_w correctly sample the $W\text{set}_{\geq 30}$ conformational space and are not restricted to some particular shapes. Let us note that RMSd are dissimilarity measures that do not necessarily respect the triangular inequality. A consequence is that the Sammon's projection does not actually reflect the word's proximity (words separated on the map can be structurally close). However, since the three point series are simultaneously projected on the same subspace, Sammon's maps can be used to qualitatively assess the similarity between the conformational sampling. We can thus conclude that OR_w are, on average, significantly more structurally stable than other words, and sample all the conformational space.

Intra-word amino-acid specificity is significantly higher for OR_w (p-value < 2×10^{-16}, cf. Table 3). The $Z_{\max}$ distributions for the three statistical word types are shown in Figure 4a. The distribution for OR_w is clearly shifted toward high values of $Z_{\max}$. OR_w are also more informative in terms of number of significant positions (p-value < 2×10^{-16}, cf. Table 3). These results must be interpreted with caution due to the restrictive condition for the interpretation of the Z-scores (see Material and Methods). However, they show that OR_w are, on average, more informative in terms of both the number of significant positions and specificity.

The coverage of sequence space by the different structural words is assessed using a procedure similar to the one used for structural space. We computed the Euclidean distances between Z-score vectors of each word pair in $W\text{set}_{\geq 30}$. The resulting average distances are given in Table 3. The Kruskal-Wallis test indicates that, in terms of amino-acid specificity, OR_w are significantly more distant one from the other (p-value $<2.2^{-16}$, cf Table 3). Sammon's map projections of the three word-types are shown in Figure 4b. We can see that OR_w cover a large region of the map, including regions not visited by NS_w and UR_w. We can conclude that OR_w are globally more distinct from each other in terms of

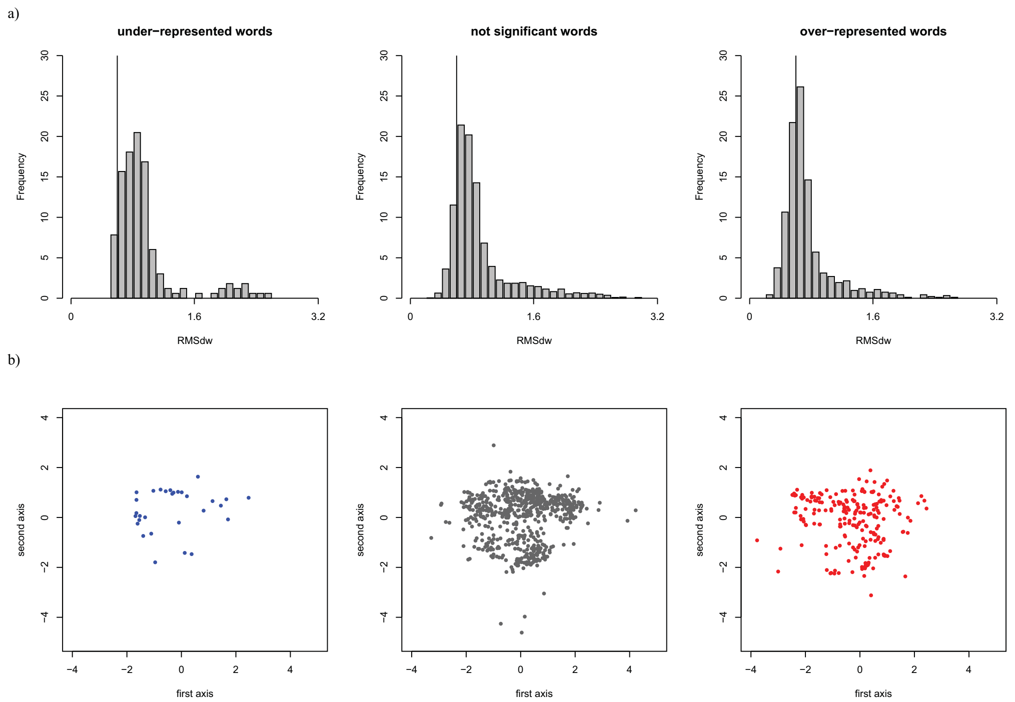

Figure 3 Structural variability of the three statistical word types of *W*set$_{\geq 30}$. a) Intra-word structural variability: distribution of the RMSd$_w$. The vertical line corresponds to a threshold of 0.6 Å b) Inter-word structural variability: Sammon's map computed from the RMSd$_{dev}$ for a sample of 890 words of *W*set$_{\geq 30}$. All the points are subjected to the same projection and plotted on distinct plots.

amino-acid sequence dependence than other words and that they sample the sequence space better than other word types.

Exceptionality and loop types

As shown in Table 1, OR_w significantly contribute to the description of long loops: OR_w cover about 40% of both short and long loops. Moreover, 58% of the loops contain at least one OR_w, and as many as 80% of long loops contain at least one OR_w. If we consider the specificity of words for a particular loop length (cf. Table 4) it can be seen that 260 OR_w are specific to long loops and 233 OR_w are specific to short loops. It means that 493 OR_w out of 930, i.e. 53% of OR_w, exhibit a significant preference for a loop-length type. This proportion should be compared to what is obtained for other words: 31% of NS_w and 28% of UR_w are significantly dependent on a particular loop length range. If we consider the flanking secondary-xstructures, the same observation can be made: 70% of OR_w versus 52% of NS_w and 45% of UR_w are specific to a particular loop type. It thus seems that OR_w exhibit stronger dependence toward the loop type than other statistical word-types.

Finally, we compared the preference of the three word-types for the eight loop-types defined by length range and flanking secondary-structures. We found that 88% of OR_w versus 72% of NS_w and 75% of UR_w exhibit a significant dependence for a particular loop type. The qualitative analysis by correspondence analysis is displayed in Figure 2, where the three statistical word types are highlighted in different colors. It can be seen that OR_w predominantly appear in outlying regions of the plot, in agreement with the KLD quantification.

Therefore, we can conclude that OR_w present higher signature in terms of structure and/or sequence and higher dependence to loop types than other words. At the same time, OR_w correctly sample all the loop-conformational space, and better cover the sequential space of protein loops. They are seen in every loop type and offer a reasonable coverage rate, with only 930 different structural motifs.

Discussion and Conclusion

In this study, we have developed an original approach for the analysis and the description of loop structures. This approach corresponds to a systematic extraction and statistical analysis of seven-residue structural motifs within loops, using a structural-alphabet simplification. Contrary to classic approaches, our method does not require either loop-structural alignment or computation of structural parameters. The structural word approach defines a structure-based clustering of all fragments, where all seven-residue fragments encoded in a similar word can be seen as a cluster. Our systematic clustering resulted in 28274 clusters, with 1 to 1633 fragments per

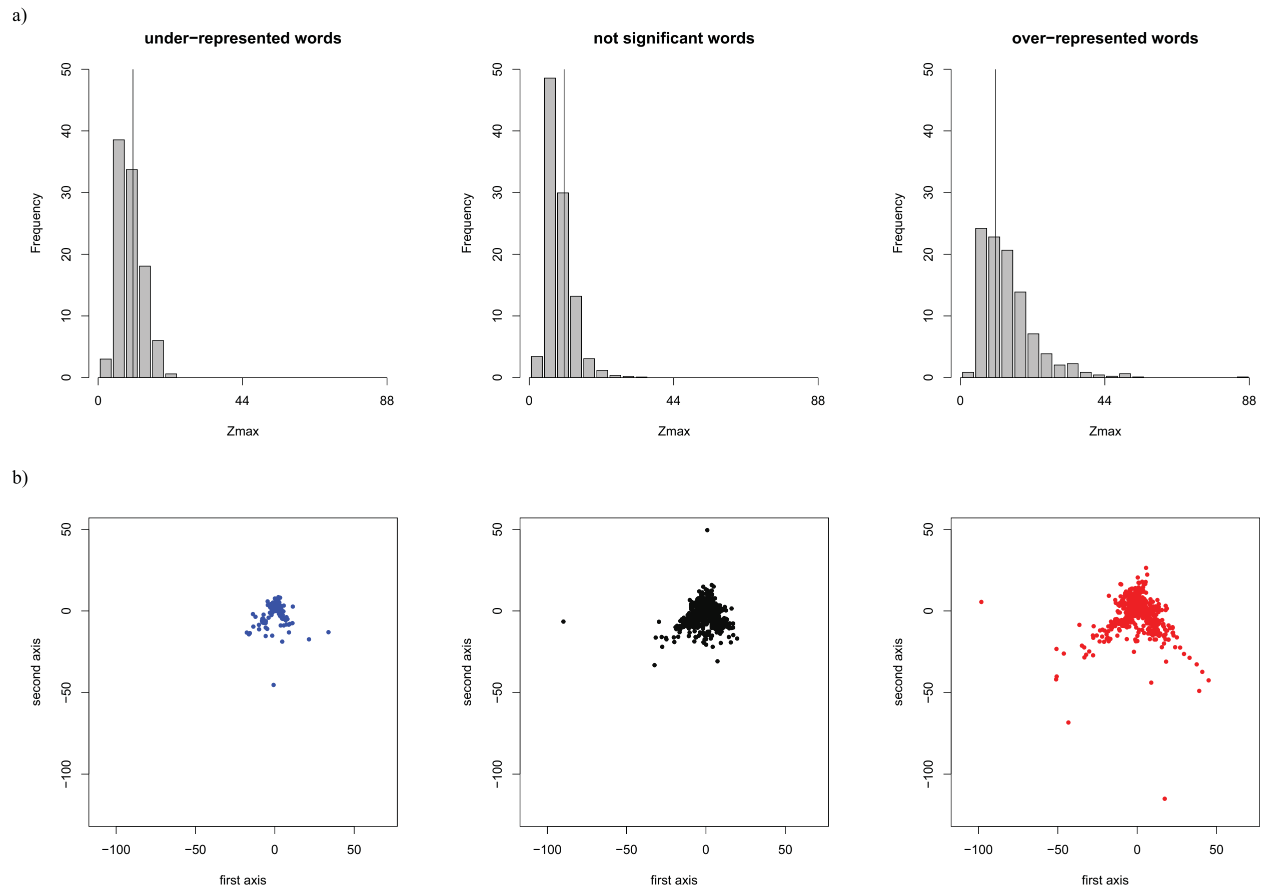

Figure 4 Sequential specificity of the three statistical word types of *W*set$_{\geq 30}$. a) Intra-word analysis: distribution of the Z_{max}. The vertical line corresponds to a threshold of 10. b) Inter-word analysis: Sammon's map computed from the Euclidean distance between *Z*-scores. All the points are subjected to the same projection and plotted on distinct plots.

cluster, and an average size equal to 15. The analysis of B-factors showed that some of the singletons are indeed associated to regions with high B-factors, which is indicative of coordinate uncertainty. It was thus legitimate to exclude them from the analysis.

In order to compute cluster properties, we chose to restrict ourselves to the 3310 clusters (= 12% of clusters) with more than 30 fragments, referred to *W*set$_{\geq 30}$. This reduction was required to have a sufficient number of fragments to compute RMSd and sequence profiles for clusters. This limited number of structural words (3310) results in a good coverage rate of the loops: 73% of loop-lengths. We additionally checked that the restriction to *W*set$_{\geq 30}$ does not result in the restriction to highly populated structural families, and that our results are stable on different data sets.

Comparison with existing approaches

An extensive comparison with already existing loop classification schemes is extremely difficult because we do not consider the same objects, and pursue different objectives. Existing classifications cluster loops according to their length [12,14,15,17,18], flanking region types [12,14,17,18,20,21], flanking region geometry [12,14,17,18] and loop geometry [17,18]. Such classifications consider full length loops and are thus inherently limited to short loops. In the present study, we cluster fixed-length structural motifs within loops, independently of their lengths or flanking regions, thus also bringing information for long loops. Consequently, it is delicate to compare our loop analysis with existing loop classifications.

Other studies have previously investigated the use of seven-residue fragments to analyze protein structures [55,56] whereas our study focuses on loop structural fragments. For this reason, the results are not directly comparable.

Other studies consisted in identifying functional patterns in whole proteins [57,58]. Such patterns, involved in protein function, are relatively rare. On the contrary, our approach considers recurrent structural motifs in loops. Alternatively, some groups have investigated the identification of 3D structural patterns linked to functions that are not necessarily made of consequent residues [59-63]. For example, Ausiello et al. (2009) [63] extracted some structural motifs from protein in different folds which recognize ligands presenting same features. In this case also, the studied objects are very different, making the comparison difficult. Another interesting analysis, MegaMotifBase, deals with structural motifs that are important for the preservation of the 3D structure in given families or superfamilies [64]. These motifs were identified using both sequence conservation and preservation of important structural features. They mainly correspond to regular secondary structures, whereas we focused our analysis on loops. For all these reasons, any comparison between our approach and already existing classifications should be regarded with caution.

Insight into loop structures

We analyzed structural and amino-acid properties of clusters, defined by structural words, using RMSd and different criteria to measure their amino-acid dependencies. We found an average intra-cluster RMSd_w equal to 0.85 Å versus 2.72 Å for the inter-cluster RMSd_{dev}, which confirms our previous results [40]. In the loop classification `ArchDB`[65] clusters grouping seven-residue loops present an average RMSd close to 1 Å. In Sander et al. [55], fragments were clustered according both to their structure and amino-acid sequence into 27 clusters with an average RMSd of 1.19Å. The most populated cluster groups α-helix fragments and probably largely contribute to the average RMSd.

Loop description by recurrent structural words permits a quantification of the loop structural redundancy: around 73% of loops are described by a limited number of accurate recurrent structural words. Thanks to the loop-structure simplification using HMM-SA, our method is the first one allowing a systematic mining of loops independently of their lengths and the study of all loops in terms of motif composition.

First, we demonstrate that the majority of the recurrent structural words have low structural variability and specific sequence signature. The simplification of loop structures using HMM-SA permits to analyze long loops. We can observe that 46% of loops are covered by words found both in short and long loops. These results show that short and long loops are composed of similar motifs. This is in agreement with the insertion/deletion process of loop evolution hypothesis made in [66]. In addition to the identification of the shared structures, our analysis provides a quantification of how the same structural words are re-used in different loops. The existence of words found in both long and short loops could allow transposing some short-loop results into the long-loop analysis and decreasing the long-loop-analysis complexity.

We observe that only one third of short (resp. long) loops are covered by words that are specific to short (resp. long) loops. Moreover, words specific to short loops have higher amino-acid specificities than other words. That means that these short loop regions (30% of short loops) are more informative in terms of sequence than other regions. Interestingly, words that are specific to long loops are structurally less variable than others meaning that a part of long loops (34%) are structurally well defined.

We also analyze the dependence between recurrent words and the loop flanking-regions. We show that around 60% of words exhibit a significant preference. Most of these words are specific to $\beta\alpha$ and $\beta\beta$ loops. These results are in agreement with classification of short loops based on flanking region information as [12,14,15,17,18,20,21] and provide an identification and quantification of the structures with a dependence on the flanking regions. Moreover, this study allows identifying and quantifying regions with no preference for flanking-region types. Indeed, 31% of loops are covered by words with no preference for a flanking-region type.

The amino-acid specificities of structural words were also assessed. We observed that 97% of recurrent words, covering 70% of loops, have amino-acid specificities. Different studies have analyzed the amino-acid preferences of loops, particularly for short loops. Kwasigroch et al. (1997) have shown that amino-acid preferences were more frequent in the core of short loops [15,16]. Other studies have focused on the amino-acid preferences of β-turns and shown that these amino-acid preferences occurred at end positions [25,67]. This study provides an identification of regions with amino-acid specificities and a new quantification of the amino-acid specificity: we found an average number of three positions with significant amino-acid preference for $W\mathrm{set}_{\geq 30}$ motifs.

Perspective in terms of loop-structure prediction

Most recurrent motifs exhibit significant amino-acid specificities: half of them display significant level of amino-acid conservation in at least four significant-positions. If we consider words with at least four significant-positions as predictable, we extract 1359 words covering 60% of the loops (on a per-structural letter basis). It is clear that this predictability index (at least four significant positions) is very basic and too optimistic. The predictability index of a word has to combine both its sequence informativity and sequence specificity. Indeed, one word can have several positions with high amino-acid preferences but close sequence from other words. Conversely, words with few informative positions can be clearly distinguishable from others in terms of sequence.

Moreover, several words can be compatible with a same seven-residue sequence, involving several candidates per amino-acid sequence. A possible strategy for loop prediction would consist in splitting the query sequence into overlapping seven-residue fragments, and identifying subset of structural words compatible in terms of sequence profile with each fragment.

The successions of compatible overlapping word candidates would then be selected using a hidden Markov model taking into account the favorable transitions between structural words. This would result in a 1D structural letter trajectory set compatible with the target loop sequence. Then, the 3D reconstruction from this set of 1D trajectories could be achieved using an energy function as in PEPfold [68]. This approach could yield a set of 3D structural conformation candidates for the target loop, in agreement with the flexibility of loops. Finally, for long loop prediction, a confidence index could be proposed for different parts of the predicted loop. Indeed, for a given loop, prediction of some regions could result in a limited number of word candidates while for other regions, the prediction could result in a large number of word candidates. This approach could be a way to decrease the complexity of long-loop prediction.

Illustrative Example of loop analysis

In Figure 5, we present an illustration of a long loop of 18 structural letters extracted from the protein structure with pdb code 3SIL, encompassing residues 120 to 140. Using the word extraction protocol, this loop was decomposed into 15 words of 4 structural letters. Among these 15 words, four words -namely UOGI, KHBB, IFFR and RPBQ- belong to $W\text{set}_{\geq 30}$. These four words are seen in both short and long loops in the data set, as illustrated in Figure 5. Structural word KHBB is over-represented, with an L_p value equal to 39.5. It is characterized by a low structural variability (RMSd_w = 0.4 Å) and strong amino-acid preference (Z_{max} = 25), with conservation of hydrophobic amino acid at position 2 and Proline at position 3. These amino-acid conservation trends are derived from the analysis of every occurrence of a particular fragment.

In this particular protein, a Lysine and a Threonine occupy positions 2 and 3 of word KHBB. This region does not appear to be particularly conserved in the multiple alignment of homologous sequences retrieved from a BLAST search in Swiss-Prot (data not shown). When aligned with sequences retrieved from a BLAST search in PDB sequences, this region exhibits three positions with equivalent residues (see alignment in Additional file 1). We attempted to further explore the functional implication of this long loop. 3SIL is a sialidase from *Salmonella typhimurium*. It corresponds to Swiss-Prot entry NANH_SALTY, and is responsible for the cleavage of terminal sialic acid from glycoproteins. There is no functional annotation in Swiss-Prot for the 120-140 region, but the catalytic and substrate-binding sites are annotated. They are highlighted in pink and blue in Figure 6. Furthermore, a structure of sialidase co-crystallized with an inhibitor is available in the PDB: structure 1DIL, with sequence identical to 3SIL. The inhibitor is thus shown in red in Figure 6. It can be seen that loop 120-140 is spatially close to functional residues and inhibitor molecules. This observation suggests that this loop could be important for the substrate stabilization, but only the observation of the enzyme co-crystallized with a substrate could confirm this hypothesis.

This example shows that some motifs extracted from loops seem to be involved in protein function. It is not surprising due to the fact loops are often involved in protein function.

Perspective of functional-motif identification

In genomic sequences, functional motifs are often characterized by particular frequencies (rare or very frequent). Therefore, the search for functional motifs is successfully guided by the search for exceptional motifs [44,45]. Inspired by this singularity, we explored the properties of structural words in proteins to see if the over- or under-representation of particular conformations can be linked to particular features. Contrary to classic methods that were primarily developed for DNA sequences, statistics are here computed by a method that takes into account the large number and short length of sequences of our data set [69]. We considered the intrinsic properties of structural words and their relationship with the statistical exceptionality status of words, classified as over-represented, under-represented, or not significant. The comparison of the three statistical word types showed that over-represented words have indeed specific properties: they are highly conserved in terms of structure or sequence and highly dependent on loop types. By setting a RMSd_w cut-off equal to 0.74 Å and a Z-max cut-off equal to 14, we found that 89% of over-represented words present either a low RMSd or a high Z_{max} or a significant dependence to a loop type defined by eight types according to the KLD criterion. This ratio is only 62% for other words. This indicates that statistical exceptionality results from a complex process combining word frequency, sequence and/or structure properties. The consideration of statistical exceptionality thus enhances the signal-to-noise ratio in protein loops. Most of the time, the relationship between local structures and protein function is not straightforward. Our findings open new perspectives to the use of over-representation in order to detect functional motifs in loops. It is the subject of an ongoing

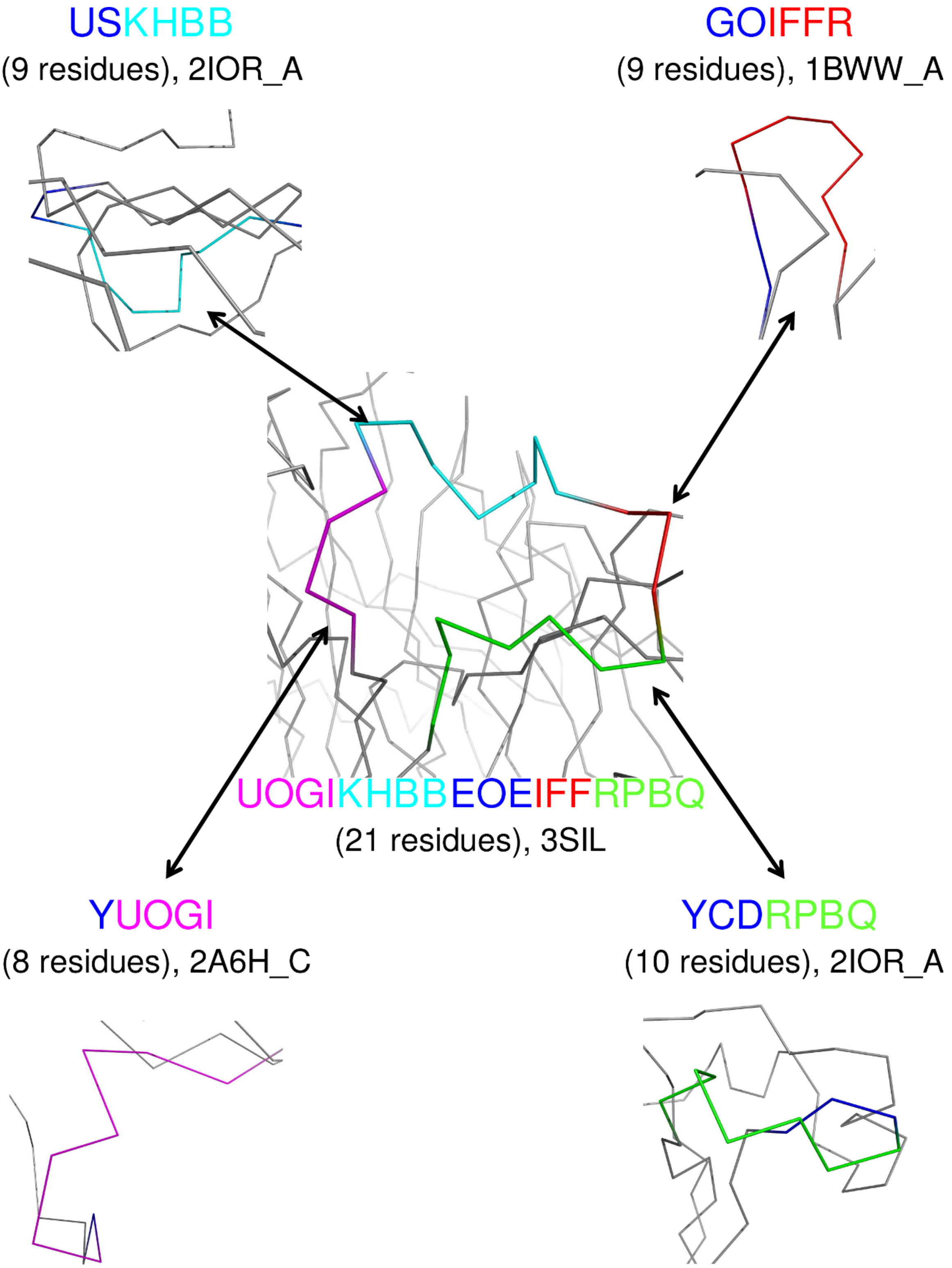

Figure 5 Recurrent words found both in long and short loops. A long loop of 18 structural letters (central figure) extracted from protein with pdb code 3SIL contains four words (`UOGI, KHBB, IFFR, RPBQ`) of *W*set$_{\geq 30}$. The protein is colored in gray, and the loop in blue except the four words `UOGI` in magenta, `KHBB` in cyan, `IFFR` in red and `RPBQ` in green. These four words are also seen in short loops in other structures. For each word, we indicate the structural letter pattern, the loop length within brackets and the pdb code of the protein structures. Structures are displayed with pymol [78].

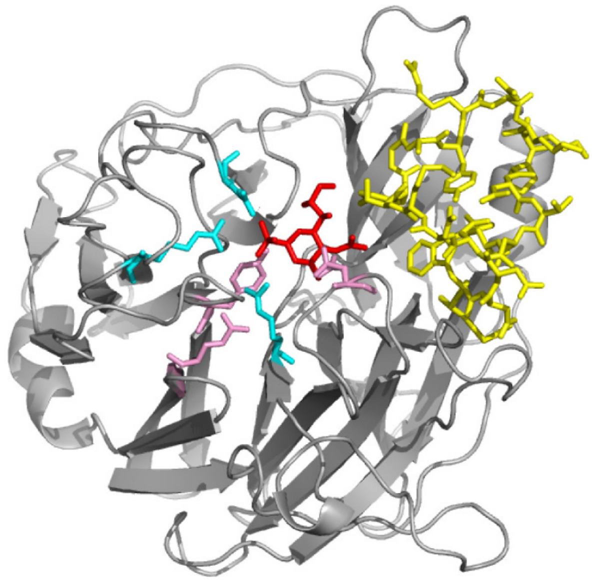

Figure 6 Functional residues of sialidase 3SIL. Catalytic and binding residues annotated in Swiss-Prot are highlighted in pink and cyan. The inhibitor (found in structure 1DIL) is highlighted in red. The long loop revealed by the structural word analysis is highlighted in yellow.

study (Regad et al, in preparation) where we suppose that functional motifs could correspond to over-represented motifs in a protein family.

Methods

Data

We used a data set of protein structures corresponding to chains presenting less than 50% of sequence pairwise identity extracted from PDB of May 2008. The data set is composed of 8186 protein chains of at least 30 residues, obtained by X-ray diffraction with a resolution better than 2.5 Å. Proteins for with missing residues or alternate conformations were removed.

Structure simplification using HMM-SA

Our structural alphabet, HMM-SA, is a library of 27 structural prototypes of four residues, called structural letters, established using a hidden Markov model [42,70]. Thanks to HMM-SA, the 3D structure of a protein backbone is simplified into a sequence of structural letters. The simplification relies on Cα positions only: each four-residue fragment of the protein structure is described by four inter-Cα distances. Consecutive four-residue fragments are overlapping on three residues resulting in one common distance. The resulting distances are the input of a hidden Markov model, and the 3D structure is translated as a sequence of 1D structural letters. This translation is made using the Viterbi algorithm [71] and takes into account both the structural similarity of the fragments with the 27 structural letters of the structural alphabet and the preferred transitions between structural letters. A protein structure of n residues is then simplified as a sequence of (n - 3) structural letters. The 27 structural letters, named [A-Z, a] are shown in Figure 1. It has been shown previously [51], that four structural-letters, [a, A, V, W], specifically describe α-helices, and five structural letters, [L, M, N, T, X], specifically describe β-strands. The remaining 18 structural letters [B, C, D, E, F, G, H, I, J, K, O, P, Q, R, S, U, Y, Z] allow accurately describing loops. Some transitions between structural letters are not possible, which results in a limited number of pathways between letters and in a limited number of short patterns of structural letters.

Extraction of structural motifs within loops

Following our previous study [40], loops are identified as series of structural letters linking simplified regular secondary structures (α-helices and β-stands) that are defined using regular expressions of structural letters. This approach permits to extract a bank of 93396 simplified loops ranging from 4 to 82 structural letters with an average length of 8.5 ± 5.5 structural letters, corresponding to an average length of 11.5 ± 8.6 residues. A loop of l structural letters corresponds to (l + 3) residues. Long loops -more than 12 residues- represent 28% of the loops in our data set. 39% of the loops are linking two β-strands, 23% are linking a β-strand to an α-helix, 22% an a-helix to a β-strand, and 16% two α-helices. The extraction of structural motifs in loops is illustrated in Figure 1. Simplified loops are split into series of overlapping words of four structural-letters, i.e., seven residues. A loop of l structural letters is then split into (l - 3) words. As we focus on structural motifs within loops, words beginning or ending with a structural letter specific to regular secondary structures [AaVWLMNTX] are excluded. This results in a global set of 28274 structural words describing all loops in the simplified structural alphabet space. The structural words thus define a partition of the structural diversity of loops, where each four-structural-letter word is a cluster of seven-residue fragments.

Loop coverage by structural words

The coverage rate of loops by a word set corresponds to the percentage of loop structural-letters covered by these words.

For example, given two loops of 11 (l_{11}) and 15 (l_{15}) structural letters and a set of recurrent 4-structural-letter words (S_w). Loop l_{11} contains two words of S_w on positions 1 to 4 and 8 to 11. As these two words are not overlapping, they cover 8 structural letters. Loop l_{15} contains three words of S_w on positions 1 to 4, 3 to 6 and 9 to 12. As the first two words are overlapping, these three words cover 10 structural letters. Thus, the coverage rate of these two loops by S_w is equal to $\frac{8+10}{11+15} = 69\%$.

This coverage rate is used in order to provide information on loop description by a set of structural words.

Structural variability of words

Intra-word

The structural variability of a structural word is measured by the geometric variability of the seven-residue fragments encoded by that word, computed using C_α Root-Mean-Square deviation (RMSd$_w$). It is obtained by computing the average RMSd$_w$ between 30 randomly selected fragments in the cluster. It is only computed for words seen more than 30 times.

Inter-word

The structural dissimilarity between two words is similarly measured by the average C_α Root-Mean-Square deviation (RMSd$_{dev}$) between 30 fragment pairs randomly selected within pairs of seven-residue fragments encoded by the two words. The word-structure-space coverage is analyzed by a Sammon's map [72] performed using the C_α RMSd$_{dev}$ dissimilarity matrix

Sequential specificity of words

Although the structural-alphabet decomposition into structural word is purely geometrical, it is still possible to analyse the sequence-to-structure dependence *a posteriori*. This is achieved using *Z*-score computation.

Intra-word

For a word *w*, we compute a *Z*-score for each of the 20 amino acids at each of the 7 positions of fragments corresponding to the word.

The *Z*-score of amino acid a, ($1 \le a \le 20$) at position l ($1 \le \ell = 7$) of a word *w*, is obtained by comparing the observed frequency of amino acid *a* at position ℓ in word *w* with its expected one:

$$Z_{a,\ell,w} = \frac{N_{a,\ell,w} - \mathbb{E}(N_{a,\ell,w})}{\sqrt{(\mathbb{V}(N_{a,\ell,w})}} \quad (1)$$

To facilitate the computation of Z-scores, we approximate the distribution of amino acid *a* in position ℓ of word *w* (corresponding to a binomial distribution $\mathscr{B}$ ($N_{a,\ell}$, $\frac{N_w}{N}$)) by a Poisson distribution $\mathcal{P}$ ($N_{a,\ell} \cdot N_w$), Where

$$\mathbb{E}(N_{a,\ell,w}) = \mathbb{V}(N_{a,\ell,w}) = \frac{N_{a,\ell} \cdot N_w}{N} \quad (2)$$

where N_w is the frequency of *w* and *N* is the total number of words in the whole data set.

To analyze the significance of a *Z*-score, the expected frequency $\mathbb{E}$ ($N_{a,\ell,\ w}$) must be greater than 5. A positive Z-score corresponds to an over-representation of the amino acid, and a negative one corresponds to an under-representation of the amino acid.

A word is thus described by a vector of 140 (7 positions × 20 amino acids) Z-scores. From these 140 Z-scores, two criteria are used to assess the amino-acid informativity of each word. The first criterion, denoted Z$_{\max}$, corresponds to the maximum *Z*-score among the 140. It measures the strongest amino-acid specificity among the 7 positions of a word. The second criterion, named nb$_{pos^*}$, $1 \le$ nb$_{pos^*} \le 7$, corresponds to the number of positions of word *w* where at least one amino acid is significant in terms of *Z*-scores. Significance cut-off is set to 4 using Bonferroni correction. It should be noted that this second criterion underestimates the sequence informativity because of the limitation introduced by the *Z*-score validity condition (only *Z*-scores with expected frequency $\mathbb{E}$ ($N_{a,\ell,\ w}$) higher than 5 can be considered for significance).

Inter-word

To check if two words have close amino-acid-sequence preferences, the Euclidean distance between their 140 *Z*-score vectors is computed [73]. The coverage of sequence specificity of words is analyzed by a Sammon's map performed using this Euclidean distance [72].

Loop type specificity of words

To study the preference of structural words for particular ℓ loop types (defined by length and/or flanking regions with ℓ, $1 \le \ell \le N_\ell$) the word distribution in different loop types is compared to the global distribution of loop types using a relative entropy measure, called the Kullback-Leibler asymmetric divergence, Kullback distance or relative entropy, denoted KLD [52]. The KLD quantifies the preference of a word w for the loop types, as:

$$KLD(w) = \sum_{\ell=1}^{N_\ell} p_{w,\ell} \cdot \log\left(\frac{p_{w,\ell}}{p_\ell}\right) \quad (3)$$

where $p_{w,\ \ell}$, denotes the relative frequency of word *w* in loop type ℓ and p_ℓ, the relative frequency of loop type ℓ among all loops. The KLD is equal to 0 if is *w* is similarly distributed in every loop type and increases with loop type dependence. The significance of KLD value is assessed by a chi-square test, since the quantity 2 × N_w × *KLD*(*w*) follows a chi-square with N_w - 1 degrees of freedom. Thus, words associated to specific loop types have significant KLD values. A correction is introduced using False Positive Rate (FPR) to take into account multiple testing. A correspondence analysis is used to visualize the main relationships between words and loop types.

Loop-word statistical exceptionality

The principle is to compare the actual frequency of a word in the data set and its expected frequency under a background reference model. A word seen significantly more (respectively less) than expected is then classified as over-represented (respectively under-represented). The expected frequency is computed using a Markov model for which the parameters are estimated from the global set of loops. This is performed using the software SPatt [74] available at http://stat.genopole.cnrs.fr/spatt, with a first order Markov chain used as reference. SPatt approach is based on the Pattern Markov Chain (PMC) notion [75]. This software has been adapted to the case of data sets with a large number of short sequences [43]. The statistical significance of the exceptionality is quantified by a p-value. To facilitate the analysis, p-values are translated into scores using equations:

$$\begin{array}{lcl} L_p & = & -\log_{10}[\mathbb{P}(N(w) > N_w)] \text{ when } w \text{ is seen more than expected} \\ L_p & = & +\log_{10}[\mathbb{P}(N(w) \leq N_w)] \text{ when } w \text{ is seen less than expected} \end{array} \quad (4)$$

where $N(w)$ is the expected frequency of the word w, and N_w its observed frequency. An over-represented word has a positive L_p value and an under-represented word has a negative L_p value. For example, an L_p equal to 21.3 means that the word is over-represented with a p-value equal to $10^{-21.3}$. A L_p equal to -17.7 means that the word is under-represented with a p-value equal to $10^{-17.7}$. The L_p threshold for statistical significance is set to 5.94, using the Bonferroni adjustment to take into account multiple tests. This permits to classify words as over-represented (L_p > 5.94), under-represented (L_p < -5.94) or not significant (-5.94 ≤ L_p ≤ 5.94).

As explained in [75], pattern significance scores tend to increase with the considered database size. This is due to the fact that a tail distribution event like the one we usually consider in pattern problems (i.e. pattern with small p-value) falls within the range of the Large Deviations theory [76,77] which means that its probability p to occur can be approximated by $p \simeq \exp(-\ell I)$ where I is a real positive rate and ℓ is the database size. As a consequence we have log $p \simeq -\ell I$ which is exactly the pattern score we consider (up to a constant multiplier). It is hence obvious that extreme pattern scores will increase in magnitude linearly with database size. If this is not a problem when we perform a pattern analysis on a single database, this bias has obviously to be corrected in order to compare results from two different databases. The correction simply consists in using one of the database as a reference and rescaling the pattern scores obtained on the second database by the appropriate ratio of sizes.

Additional file 1: Supplementary. This file is a pdf file. It contains different information about: • Extraction of words of different lengths. • Comparison of the loop length distribution in loops containing all words and loops containing only words seen 30 times. • Coverage of SCOP superfamilies by recurrent words. • Correlation between sequence specificity ($Z_{\max}$) and structure variability (RMSd$_w$) for all words in $W\text{set}_{\geq 30}$. • Exceptionality score L_p versus frequency for the 28274 words of the data set. • Robustness of the word statistical analysis on different data sets. • ClustalW of 3SIL sequence (P29768) and homologous sequences from UniProt
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-75-S1.PDF]

Acknowledgements

The authors want to express their appreciation and wish to acknowledge Pr. Philippe Deureumaux, Pr. Gilles Labesse, Dr. Jacques Chomilier, Dr. Joel Pothier for helpful discussions, Dr. Christelle Reynès and Dr. Joel Pothier for their critical reading of the manuscript, and Dr. Gaelle Debret for her help. The authors also wish to acknowledge the anonymous reviewers for their thoughtful remarks that helped to improved the manuscript. LR had a grant from the Ministère de la Recherche. JM had a grant from INRA.

Author details

[1]MTi, Inserm UMR-S 973, Université Paris Diderot- Paris 7, Paris, F-75205 Cedex 13, France. [2]Unite Mathématiques Informatique et Génome UR1077, INRA, Jouy-en-Josas, F-78350, France. [3]Université Lyon 1, IFR 128, CNRS, UMR 5086, IBCP, Institut de Biologie et Chimie des Protéines, Lyon, F-69367, France. [4]MAP5, UMR CNRS 8145, Université Paris-Descartes,Paris, F-75006, France.

Authors' contributions

LR, JM, and ACC conceptualized the project. LR developed the software, performed the experiments and drafted the paper. NG developed and adapted the software SPatt. LR, JM and ACC analyzed the experimental results. LR, JM and ACC contributed to writing the paper. All authors read and approved the final manuscript.

Received: 7 July 2009
Accepted: 4 February 2010 Published: 4 February 2010

References

1. Fetrow JS: **Omega loops: nonregular secondary structures significant in protein function and stability.** *FASEB J* 1995, **9**:708-717.
2. Johnson LN, Lowe ED, Noble ME, Owen DJ: **The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases.** *FEBS Lett* 1998, **430**:1-11.
3. Bernstein LS, Ramineni S, Hague C, Cladman W, Chidiac P, Levey AI, Hepler JR: **RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate Gq/11alpha signaling.** *J Biol Chem* 2004, **279**:21248-21256.
4. Kiss C, Fisher H, Pesavento E, Dai M, Valero R, Ovecka M, Nolan R, Phipps ML, Velappan N, Chasteen L, Martinez JS, Waldo GS, Pavlik P, Bradbury AR: **Antibody binding loop insertions as diversity elements.** *Nucl Acids Res* 2006, **34**:132-146.
5. Saraste M, Sibbald PR, Wittinghofer A: **The P-loop: a common motif in ATP- and GTP-binding proteins.** *Trends Biochem Sci* 1990, **15**:430-434.
6. Via A, Ferre F, Brannetti B, Valencia A, Helmer-Citterich M: **Three-dimensional view of the surface motif associated with the P-loop structure: cis and trans cases of convergent evolution.** *J Mol Biol* 2000, **303(4)**:455-465.
7. Stuart D, Acharya K, Walker N, Smith S, Lewis M, Phillips D: **Lactalbumin possesses a novel calcium binding loop.** *Nature* 1986, **324**:84-87.

8. Golovin A, Henrick K: **MSDmotif: exploring protein sites and motifs.** *BMC Bioinformatics* 2008, **9**:312-312.
9. Benner SA, Gerloff D: **Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases.** *Adv Enzyme Regul* 1991, **31**:121-181.
10. Benner SA, Cohen MA, Gonnet GH: **Empirical and structural models for insertions and deletions in the divergent evolution of proteins.** *J Mol Biol* 1993, **229**:1065-1082.
11. Panchenko AR, Madej T: **Structural similarity of loops in protein families: toward the understanding of protein evolution.** *BMC Evol Biol* 2005, **5**:10.
12. Donate LE, Rufino SD, Canard LH, Blundell TL: **Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction.** *Protein Sci* 1996, **5(12)**:2600-2616.
13. Rufino SD, Donate LE, Canard LH, Blundell TL: **Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling.** *J Mol Biol* 1997, **267**:352-367.
14. Burke DF, Deane CM, Blundell TL: **Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure.** *Bioinformatics* 2000, **16**:513-19.
15. Kwasigroch JM, Chomilier J, Mornon JP: **A global taxonomy of loops in globular proteins.** *J Mol Biol* 1996, **259**:855-872.
16. Wojcik J, Mornon JP, Chomilier J: **New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification.** *J Mol Biol* 1999, **289**:1469-1490.
17. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ: **An automated classification of the structure of protein loops.** *J Mol Biol* 1997, **266**:814-830.
18. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJE, Oliva B: **ArchDB: automated protein loop classification as a tool for structural genomics.** *Nucl Acids Res* 2004, , 32 Database: 185-188.
19. Fernandez-Fuentes N, Hermoso A, Espadaler J, Querol E, Aviles FX, Oliva B: **Classification of common functional loops of kinase super-families.** *Proteins* 2004, **56(3)**:539-555.
20. Li W, Liu Z, Lai L: **Protein loops on structurally similar scaffolds: database and conformational analysis.** *Biopolymers* 1999, **49**:481.
21. Li W, Liang S, Wang R, Lai L, Han Y: **Exploring the conformational diversity of loops on conserved frameworks.** *Protein Eng* 1999, **12(12)**:1075-1086.
22. Venkatachalam CM: **Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units.** *Biopolymers* 1968, 1425-1436, Biopolymers.
23. Lewis PN, Momany FA, Scheraga HA: **Chain reversals in proteins.** *Bioch Biophys Acta* 1973, **303**:211-229.
24. Richardson JS: **The anatomy and taxonomy of protein structure.** *Adv Protein Chem* 1981, **34**:167-339.
25. Hutchinson EG, Thornton JM: **A revised set of potentials for β-turn formation in proteins.** *Protein Sci* 1994, **3**:2207-2216.
26. Sibanda BL, Thornton JM: **Beta-hairpin families in globular proteins.** *Nature* 1985, **316**:170-174.
27. Milner-White EJ, Poet R: **Four classes of beta-hairpins in proteins.** *Biochem J* 1986, **240**:289-292.
28. Sibanda BL, Blundell TL, Thornton JM: **Conformation of beta-hairpins in protein structures systematic classification with applications to modelling by homology, electron density fitting and protein engineering.** *J Mol Biol* 1989, **206**:759-777.
29. Sibanda BL, Thornton JM: **Conformation of β hairpins in protein structures: classification and diversity in homologous structures.** *Methods Enzymol* 1991, **202**:59-82.
30. Efimov A: **Structure of coiled β - β hairpins and β - β corners.** *FEBS* 1991, **284**:288-292.
31. Rice PA, Goldman A, Steitz TA: **A helix-turn-strand structural motif common in alpha-beta proteins.** *Proteins* 1990, **8(4)**:334-340.
32. Leszczynski JF, Rose GD: **Loops in globular proteins: a novel category of secondary structure.** *Science* 1986, **234**:849-855.
33. Kabsch W, Sander C: **Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22**:2577-637.
34. Matthews BW: **The gamma turn. Evidence for a new folded conformation in proteins.** *Macromolecules* 1972, **5**:818-819.
35. Rose GD, Gierasch LM, Smith JA: **Turns in peptides and proteins.** *Adv Protein Chem* 1985, **37**:1-109.
36. Milner-White EJ, Ross BM, Ismail R, Belhadj-Mostefa K, Poet R: **One type of gamma-turn, rather than the other gives rise to chain reversal in proteins.** *J Mol Biol* 1988, **204**:777-782.
37. Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C, Saviano M: **Discovering protein secondary structures: classification and description of isolated α-turns.** *Biopolymers* 1996, **38**:705-721.
38. Chou KC: **Prediction of tight turns and their types in proteins.** *Anal Biochem* 2000, **286**:1-16.
39. Leader D, Milner-White E: **Motivated proteins: a web application for studying small three-dimensional protein motifs.** *BMC Bioinformatics* 2009, **10**:60-60.
40. Regad L, Martin J, Camproux AC: **Identification of non Random Motifs in Loops Using a Structural Alphabet.** *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational, Toronto* 2006, 92-100.
41. Kolodny R, Koehl P, Guibas L, Levitt M: **Small libraries of protein fragments model native protein structures accurately.** *J Mol Biol* 2002, **323**:297-307.
42. Camproux AC, Gautier R, Tufféry T: **A hidden Markov model derivated structural alphabet for proteins.** *J Mol Biol* 2004, **339**:561-605.
43. Nuel G, Regad L, Martin J, Camproux AC: **Exact distribution of pattern in a set of random sequences generated by a Markov source: application to biological data.** *Algo Mol Biol* 2010, **5**:15.
44. Leung MY, Marsh GM, Speed TP: **Over- and underrepresentation of short DNA words in herpesvirus genomes.** *J Comput Biol* 1997, **3**:345-360.
45. Rocha E, Viari A, Danchin A: **Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons.** *Nucl Acids Res* 1998, **26**:2971-2980.
46. Karlin S, Burge C, Campbell AM: **Statistical analyses of counts and distributions of restriction sites in DNA sequences.** *Nucl Acids Res* 1992, **20**:1363-1370.
47. Sourice S, Biaudet V, El Karoui M, Ehrlich S, Gruss A: **Identification of the Chi site of Haemophilus influenzae as several sequences related to Escherichia coli Chi site.** *Mol Microbiol* 1998, **27**:1021-1029.
48. van Helden J, Olmo M, Perez-Ortin JE: **Statistical analysis of yeast genomic downstream sequences revels putative polyadenylation signals.** *Nucl Acids Res* 2000, **28**:1000-1010.
49. Mönnigmann M, Floudas C: **Protein loop structure prediction with flexible stem geometries.** *Proteins* 2005, **61(4)**:748-62.
50. Bourne PE, Weissig H: **Structural Bioinformatics (Methods of Biochemical Analysis).** Wiley-Liss 2003 chap. Structure Quality Assurance**44.**
51. Camproux AC, Tufféry P: **Hidden Markov Model-derived structural alphabet for proteins : the learning of protein local shapes captures sequences specificity.** *Biochim Biophys Acta* 2005, **1724**:394-403.
52. Kullback S, Leibler R: **On information and sufficiency.** *Annals of Mathematics and Statistics* 1951, **22**:79-86.
53. Fuchs P, Alix JF, Alain JP: **High accuracy prediction of beta-turns and their types using propensities and multiple alignments.** *Proteins* 2005, **59**:828-839.
54. Hollander M, Wolfe DA: **Nonparametric statistical inference.** New York: John Wiley and Son 1973.
55. Sander O, Ingolf S, Lengauer T: **Local protein structure prediction using discriminative models.** *BMC Bioinformatics* 2006, **7**:14-26.
56. Hunter CG, Subramaniam S: **Protein fragment clustering and canonical local shapes.** *Proteins* 2003, **50**:580-588.
57. Espadaler J, Querol E, Aviles FX, Oliva B: **Identification of function-associated loop motifs and application to protein function prediction.** *Bioinformatics* 2006, **22**:2237-2243.
58. Kim S, Wang Z, Dalkilie M: **iGibbs: Improving Gibbs Motif Sampler for proteins by sequence clustering and iterative pattern sampling.** *Proteins* 2007, **66**:671-681.
59. Torrance JW, Bartlett GJ, Porter CT, Thornton JM: **Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families.** *J Mol Biol* 2005, **347**:565-581.
60. Polacco BJ, Babbitt PC: **Automated discovery of 3D motifs for protein function annotation.** *Bioinformatics* 2006, **22**:723-730.

61. Sacan A, Ozturk O, Ferhatosmanoglu H, Wang Y: **LFM-Pro: a tool for detecting significant local structural sites in proteins.** *Bioinformatics* 2007, **23**:709-716.
62. Ausiello G, Gherardini P, Marcatili P, Tramontano A, Via A, Helmer-Citterich M: **FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures.** *BMC Bioinformatics* 2008, **9**:S2.
63. Ausiello G, Gherardini P, Gatti E, Incani o, Helmer-Citterich M: **Structural motifs recurring in different folds recognize the same ligand fragments.** *BMC Bioinformatics* 2009, **10**:182-191.
64. Pugalenthi G, Suganthan PN, Sowdhamini R, Chakrabarti S: **MegaMotifBase: a database of structural motifs in protein families and superfamilies.** *Nucleic Acids Res* 2008, **36**:D218-221.
65. Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJE, Oliva B: **Prediction of conformation and geometry of loops in globular proteins; Testing ArchDB, a structural classification of loops.** *Proteins* 2005, **60**:746-757.
66. Panchenko AR, Madej T: **Analysis of Protein Homology by Assessing the Dis(similarity) in Protein loop regions.** *Proteins* 2004, **57**:539-547.
67. Colloc'h N, Cohen F: **Beta-breakers: an aperiodic secondary structure.** *J Mol Biol* 1991, **221(2)**:603-13.
68. Maupetit J, Derreumaux P, Tuffery P: **PEP-FOLD: an online resource for de novo peptide structure prediction.** *Nucleic Acids Res* 2009, , **37 Web Server:** W498-503.
69. Martin J, Regad L, Camproux AC, Nuel G: **Pattern statistics in set of biological short sequences.** *ASMDA Proceedings* 2007, 1-10.
70. Camproux AC, Tufféry P, Chevrolat JP, Boisvieux J, Hazout S: **Hidden Markov model approach for identifying the modular framework of the protein backbone.** *Protein Eng* 1999, **12**:1063-1073.
71. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition.** *Proceedings of the IEEE* 1989, **77**:257-286.
72. Sammon JW: **A non-linear mapping for data structure analysis.** *IEEE Trans Comput* 1969, **C-18**:401-409.
73. Martin J, de Brevern AG, Camproux AC: **In silico local structure approach: a case study on Outer Membrane Proteins.** *Proteins* 2007, **71**:92-109.
74. Nuel G: **S-SPatt: simple statistics for patterns on Markov chains.** *Bioinformatics* 2005, **21**:3051-3052.
75. Nuel G: **Numerical solutions for Patterns Statistics on Markov chains.** *Statistical Applications in Genetics and Molecular Biology* 2006, **5**:26.
76. Dembo A, Zeitouni O: **Large deviations techniques and applications.** Springer 1998.
77. den Hollander F: **Large deviations.** American mathematical society, Providence 2000.
78. DeLano WL: **The PyMOL Molecular Graphics System.** 2002http://www.pymol.org.

doi:10.1186/1471-2105-11-75
Cite this article as: Regad *et al.*: Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics* 2010 **11**:75.