Evaluating Protein Motif Significance Measures: A Case Study on Prosite Patterns

Pedro Gabriel Ferreira
Department of Informatics,
University of Minho,
Portugal
pedrogabriel@di.uminho.pt

Paulo J. Azevedo
Department of Informatics,
University of Minho,
Portugal
pja@di.uminho.pt

***Abstract*— The existence of preserved subsequences in a set of related protein sequences suggests that they might play a structural and functional role in protein's mechanisms. Due to its exploratory approach, the mining process tends to deliver a large number of motifs. Therefore it is critical to release methods that identify relevant significant motifs.**

Many measures of interest and significance have been proposed. However, since motifs have a wide range of applications, how to choose the appropriate significance measures is application dependent. Some measures show consistent results being highly correlated, while others show disagreements. In this paper we review existent measures and study their behavior in order to assist the selection of the most appropriate set of measures. An experimental evaluation of the measures for high quality patterns from the Prosite database is presented.

I. Introduction

The mining of *sequence patterns* or *motifs* is one of the most important tasks in protein sequence analysis and continues to be an active topic of research. The large number of works and algorithms that can be found in literature sustain this claim. Sequence mining consists in the task of analyzing a set of possible related sequences and detecting subsequences, also called *motifs*, that occur a significant number of times among those sequences. The motif overrepresentation can be explained by the existence of segments that have been preserved through natural evolution of the protein. This may suggests that those subsequences play a structural and functional role in the protein's mechanisms [32], [7]. Different types of motifs representation have been proposed and two main classes can be distinguished: *probabilistic* and *deterministic*. Probabilistic motifs consist of a model that simulates the sequences or part of the sequences under consideration. When an input sequence is provided, a probability of being matched by the motif is yielded. Position Weight Matrices (PWM) and Hidden Markov Models (HMMs) are examples of probabilistic motifs. Deterministic motifs are commonly expressed through means of an enhanced regular expression syntax, either matching or not the input sequences.

A critical aspect that is raised during motif analysis is that the mining process tends to report a large number of motifs. This can be blamed to the algorithm characteristics, the database properties or even the user parameter values. Not all these motifs are particularly interesting and most of them certainly arise by chance. It is therefore crucial to propose pruning methods to discriminate the relevant and significant motifs.

The definition of significant motif is by itself an interesting problem. One possible solution to asses the significance of motifs is to delegate this decision on a biologist. This expert would analyze the target proteins and decide which motifs have biological significance. Since this approach is not feasible in practice, an alternative is to automatically evaluate the motifs according to their statistical or informative importance. As pointed in [26] by Stolovitzy and Califano, although the statistical significance can be neither necessary nor sufficient for biological significance, it provides a starting point for this analysis.

Additionally to the task of supporting a better understanding of the protein's structure and function, motifs have also a wide-range of other applications. They can be used to perform clustering [30], family classification [6], [7], [11], [15], [20], [14], [8], discovery of sub-families in large protein families [1], sequence annotation, gene expression analysis [17] and the study and discovery of homology relations. The selection of the appropriate measures for a specific problem depends on how well they adjust to the problem.

In the literature, many measures of interest and significance have been proposed. Usually, for each proposed motif mining algorithm a different measure is also proposed. How to choose the most appropriate significance measure is still an open question. As far as we know, a thorough survey on comparing different metrics for motif significance is still missing. Such study can bring significant improvements to the field of protein sequence analysis. For instance, the unsupervised mining of massive protein datasets (like comprehensive protein sequence databases e.g. SwissProt [12]) is not yet possible. This can be due to the limitations of the existent algorithms which are not yet capable of efficiently mining such amount of data. However, the inexistence of measures that objectively evaluate the biological significance of newly discovered motifs also contribute to preclude this goal.

Different measures have different properties, thus the best solution for a particular problem will most probably include several measures and not be reduced to just choosing just one. Given that some of these measures will show consistent and similar results, it is important to study how they interrelate. This will permit the identification of mutually different measures and avoid biased evaluations. It is our aim to identify such relations among the studied measures. The contributions of this paper can be summarized as follows:

- Discuss the need of making a comprehensive evaluation of different motif significance measures.
- Survey different types of measures presented in the bioinformatics, data mining, statistics and machine learning literature and provide a full characterization.
- Evaluate the measures on a set of well defined motifs. Study the behavior of these measure on different motifs (intra-measure information) and how they inter-relate (inter-measure information).

II. Preliminaries

A. Evaluating Deterministic Motifs

Deterministic motifs make use of widely known regular expression syntax, being easily understandable by humans. These motifs can be divided in two types: *fixed-length* and *extensible-length*. Fixed-length motifs (a.k.a (l, d)-motifs [23], [10]) consist of motifs of a fixed size of l symbols where d possible symbols may have a mismatch with the matched subsequences in a database. Extensible-length motifs have an arbitrary length. In the enhanced regular expression syntax by which they are expressed, each symbol is generically called an *event* and the distance between consecutive events as *gaps*. Consider the definition of extensible-length pattern as:
$A_1 - x(p_1, q_1) - A_2 - x(p_2, q_2) - \ldots - A_n$
, where A_i is a sequence of consecutive amino-acids and $-x(p_i, q_i)-$ gaps greater than p_i and smaller than q_i. Two types of extensible-length motifs can be distinguished:

- **Rigid Gap Motifs** only contain gaps with a fixed length, i.e. $p_i = q_i, \forall i$. The symbol "." is a wild-card symbol used to denote a gap of size one and it matches any symbol of the alphabet. Ex: $MN..A.CA$
- **Flexible Gap Motifs** allow a variable number of gaps between events of the sequence, i.e. $p_i \leq q_i, \forall i$.
Ex: AN-x(1,3)-C-x(4,6)-D.

Deterministic motifs are typically mined through combinatorial algorithms that perform an exhaustive traversal of the search space and output motifs based on the support metric. This metric requires that a motif, in order to be reported, occurs in a number of sequences equal or greater than a user pre-defined threshold (see to [21] for a comprehensive overview). Typically, the assessment of the motifs significance is done as post-processing step. In this context, two important facts justify the critical need to evaluate significance measures. The first is to provide means for an early pruning of irrelevant motifs. As a result of the combinatorial nature of the mining algorithms, the number of potentially candidates of deterministic motifs can easily grow exponentially. The second refers to the fact that the over-representation implied by the minimum support threshold does not always implies the significance of the motif.

B. The Prosite Database

Today, there are a significant number of motifs repositories freely available at the Internet. Examples of well established and reliable sequence motif databases are: *Prosite*, *PRINTS*, *BLOCKS*, *InterPro* or *eMotif*(see [16] for an overview). From the mentioned databases, Prosite deserves a special attention in the context of our work. Prosite [4] is the oldest and best known sequence motif database. It is a semi-manually annotated database. The sequence motifs are characterized by having a high biological significance. They provide a strong indication of a region in the protein with an important role. A family of protein sequences is then described by one or more motifs. The key aspects of the Prosite motifs are: its capability to identify functional or structural regions in the proteins and its use as a tool to distinguish family members. This database is considered a standard (due to the high quality of its motifs). New algorithms and methods tend to use this database as a benchmark test. As an example of a motif, we examine the Prosite entry ps00017. It reports the ATP/GTP-binding site motif A also known as P-loop. This motif appears in a considerable number of proteins that bind ATP or GTP. It is a motif with a high probability of occurrence. A scan to Swiss-Prot (release 49.1) shows that this motif has 17861 hits in 16550 different sequences. The pattern has the format: [AG] - x(4) - G - K - [ST].

III. Significance Measures

As introduced by Brazma et al. in [9], a significance measure can be defined as function in the form: $f(m, C) \rightarrow \mathbb{R}$. m stands for the motif being evaluated and C a set of possibly related proteins sequences. This function returns a real valued score that expresses how relevant or significant is m with respect to C. These scores may provide hints to biological or statistical relevant motifs.

The set of sequences C under which m is being compared usually correspond to a part or the totality of a family of proteins and is called *target family*. The set of remaining sequences in the universe of all sequences U is denoted as $\overline{C}$, where $U = C + \overline{C}$. The size of each set of sequences is denoted as $|C|$ and $|\overline{C}|$, respectively. We now distinguish four possible cases of a motif m matching a sequence of C:

- *True Positive* (T_P): a sequence that belongs to the target family and matches the motif.
- *True Negative* (T_N): a sequence that does not belong to the target family and does not match the motif.
- *False Negative* (F_N): a sequence that belongs to the target family and does not match the motif.
- *False Positive* (F_P): a sequence that does not belong to the target family and matches the motif.

In [24] Sagot suggests that motifs can be evaluated according to the following approaches :

- Probability of matching a random sequence;
- Sensitivity/Specificity;
- Information content;
- Minimum Description Length;

Since this categorization does not include all the possible measures nor distinguishes the type of information provided, in this work we will adopt a different categorization. Three categories of measures are proposed:

1) *Class-based* measures are calculated based on the information of the pattern in relation to the target and complement protein classes/families.
2) *Theoretic-Information* measures are based solely on theoretic models like probabilistic or entropy models. In this case the measure calculation is self-contained, i.e. the necessary information is found in the motif itself.
3) *Mixed* measures use both theoretic and class information.

A. Class-based Measures

The ideal motif is one that matches all the sequences of the target family and no other sequence outside this family. These motifs are known as *signatures* [9] and are a perfect tool to distinguish sequences from different families. In the bioinformatics context, the measures most widely used to express the quality of the patterns are: *sensitivity*, *specificity* and *positive predicted value* (PPV) (see Table I). *Sensitivity*, also called recall, measures the proportion of sequences of the target family correctly matched by the motif. *Specificity* measures the proportion of sequences outside the target family that are not matched by the motif. The *PPV*, also called precision, measures the proportion of sequences that are covered by the motif and that belong to the target family. An ideal motif is one with 100% of sensitivity and 100% of PPV. These three measures yield a positive rank of motifs, i. e. their score is proportional to the rank. For comparison purposes we will introduce a negative rank measure: *False positive rate - Fpr*. This measure returns the proportion of negative instances that were incorrectly reported as being positive. In this case, the greater the score the worst the quality of the motif. Motifs can be ranked according to one or all of these measures. When a unique value is required to score a motif, a combination of these measures can be used. The *F-Measure* (F) [27] and the Pearson *Correlation* (C) [9], [5](also knows as Matthews Correlation Coefficient, for its application in secondary structure prediction [34]) are examples of such composed measures. As a last example of a class-based measure we refer the *Discrimination power* (Dp) [7]. This measure is particularly useful as a filter since Dp is proportionally associated to selectiveness. A characteristic of the class-based measures is that they do not rely on the motif in order to be calculated. Hence, they can be applied to any type of deterministic motif. We do not review all the possible class-based measures as many other measures covering different aspects of the pattern quality can be found. Thus, we only focus on the most intuitive and widely known measures.

B. Theoretic-Information Measures

When analyzing the probabilistic aspects of the protein sequences, it is generally assumed that sequences are generated by an independent identically distributed (i.i.d.) process. Typically, the Bernoulli model is used. Therefore, the occurrence of a motif m in a given sequence is assumed to be an i.i.d. process [3]. In practice, this means that sequences are considered to be independent and the occurrence of the amino-acids independent events. Although this argument is not always totally true (sequences are believed to be biologically related) it provides a simplification which is a good approximation to actual verified values [22]. The probability P of a motif M, in the form $A_1 - x(p_1, q_1) - A_2 - x(p_2, q_2) - \ldots - A_n$, can be calculated according to equation 1, where A_i is a subsequence of amino-acids.

$$P(M) = P(A_1) \times P(-x(p_1, q_1)-) \times P(A_2) \times P(-x(p_2, q_2)-) \times \ldots P(A_n) \quad (1)$$

Since $P(.) = 1$, then $P(-x(p,q)-) = 1$ and $P(A_i) = \prod_{a_j \in A_i} P(a_j)$. We consider that the probability of an amino-acid a_j, $P(a_j)$, is given by its frequency of occurrence at the Swiss-Prot database [12]. If ambiguous positions occurs in subsequence A_i then its probability is given by equation 2.

$$P(A_i) = \prod_{a_j \in A_i} (\sum_{k=1}^{|A_i|} P(a_{jk})) \quad (2)$$

Where a_{jk} stands for the *k-th* amino-acid in position j of the subsequence A_i. For instance, the probability of the subsequence $A - [GC] - .. - V$ is given by $0.0783 \times (0.0693 + 0.0152) \times 1 \times 1 \times 0.0671 = 4.44 \times 10^{-4}$.

$Support(M)$ is the number of times that a motif M occurs in different sequences of the database. $Support(M \cup C)$ corresponds to the number of sequences in family C where M occurs.

Theoretic-information measures quantify the degree of information encoded in a motif. We provide examples of three of these measures. The *Information Gain* (IG) [29], [28] is used to measure the amount of accumulated information of a motif in relation to an amino-acid sequence. In this equation (see table I) the information content I measures how likely a pattern is to occur and the $Support(M) - 1$ gives the recurrence of the motif M in the database.

The Minimum Description Length - MDL principle - applied in [22], [1] is also an information-theoretic measure and can be made equivalent to the IG measures. The MDL is used to score the motifs and to measure the fitness of these motifs with respect to the input sequences. Assuming the hypothetical transmission of sequences, the idea is to measure how much can be saved in this transmission, if one knows about the presence of the motif. In [22], it is demonstrated that $K \times log_2 P(M)$ is the saving obtained from a motif M over K covered sequences, being equivalent to the IG formula.

The *Log-Odds* (L) measure provides the degree of surprise of a pattern. It compares its probability of occurrence with the expected probability of occurrence according to the background distribution. The equation presented in table I is a variation of the log-odds equation introduced in [19], that was first proposed to measure the significance of probabilistic patterns. Both IG and L measures can be applied to all types of deterministic patterns. The *Pratt* (PR) measure was introduced by Jonassen et al. in [18] to rank the extensible gap motifs obtained from the Pratt algorithm. In this measure, a penalty is given when gaps occur.

As an additional measure we propose the Z-score measure. Although it is essentially a statical measure, it was included in this group of measures as it can be calculated based solely on the support, the motif information and in the number of amino-acids of the database. This measure can be used to filter out irrelevant motifs by selecting only those for which the actual number of occurrences considerably exceeds its expected number. This criteria is based on the following biological motivation: if a motif occurs more than it is expected to occur by chance then it should have a biological interest.

In [3], [26], motifs are statistically evaluated according to a *Z-Score* function. For a motif M, the Z-score is provided by equation 3 (see also table I):

$$Zscore(M) = \frac{Support(M) - E(M)}{N(M)} \tag{3}$$

In equation 3, $Support(M)$ denotes the actual number of occurrences (support) and $E(M)$ the expected number of occurrences of M, which is calculated by the product of the total number of amino-acids found in the database by the probability of M. $N(M)$ is an expected value of some function of M, in this case the square root of the expected variance. It was generally verified that statistically relevant motifs, discriminated through the Z-score function, match functionally important regions of the proteins [3], [26]. Another important conclusion obtained from [3] is that for over-represented motifs, the non-maximal motifs (which are contained on other motifs) have a lower degree of surprise than the maximal ones. This result yields a clever mechanism to prune motifs just before their significance is computed. The over-representation approach provided by means of a support constraint and the under-representation approach provided by statistical tools like the Z-score, is complementary on the task of automatically retrieving significant motifs from a database.

C. Mixed Measures

As examples of mixed measures that use information-theoretic and class-based features to determine the significance of a pattern, we selected two that are popular in the Machine Learning and Data Mining communities. These are the *J-Measure* [25] and the *Mutual Information* (I-measure) which is derived from the Shannon's entropy theory [2], [35].

For a class space $Q = \{C, \overline{C}\}$, the component $H(Q)$ of the I measure (see table I) provides the degree of information encoded by Q. Given a motif M, component $H(Q|M)$ measures the amount of uncertainty remaining about Q after M is known. The difference $H(Q) - H(Q|M)$ provides the expected information gain about Q upon knowing M.

The J-measure is the product of two components. The first component, $P(M)$, provides the probability of the motif occurrence, which can be interpreted as a measure of simplicity. Since we are considering a target class C and its complement $\overline{C}$, the second component $j(C; M)$ measures the goodness-of-fit of M with relation to class C. This is also called *cross-entropy* [36].

In addition, we redefine the IG measure to account for the distribution of motifs along the protein families, leading to the proposal of a measure called *(S)urprise-Measure*. The S measure combines the information gain, I, of the motif M with the conditional probability of matching a sequence s from the target class C given the motif M. This probability is given by the relative occurrence of M in C, $\frac{Support(M \cup C)}{Support(M)}$, which corresponds to the positive predicted value of M. This measure express the amount of information provided by the pattern and its quality as a class descriptor.

These three measures can be easily calculated in all types of deterministic motifs. In general, one can interpret these mixed measures as a tool to quantify the uncertainty reduction of a sequence *S* belonging to the class *C*, given that *S* contains the motif *M*.

Table II contains equations to support a better understanding of the ones from Table I.

Formula	Range
$P(C) = \frac{T_P + F_N}{T_P + F_N + F_P + T_N}$	[0,1]
$P(C\|M) = \frac{T_P}{T_P + F_P}$	[0,1]
$\frac{P(C\|M)}{P(C)} = \frac{T_P \times (T_P + F_N + F_P + T_N)}{(T_P + F_P) \times (T_P + F_N)}$	[0,1]
$\frac{1 - P(C\|M)}{1 - P(C)} = \frac{F_P \times (T_P + F_N + F_P + T_N)}{(T_P + F_P) \times (T_N + F_P)}$	[0,1]

TABLE II

AUXILIARY FORMULAS.

IV. EVALUATION

In this section we describe the set of experiments that were performed to evaluate the relations among the different measures. For this purpose only flexible-length motifs were evaluated. The file *prosite.dat* that corresponds to the Prosite database (available by FTP) was analyzed. It corresponds to the Release 19.20 (Feb-2006). This release contains 1929 entries, where 1330 are regular expression patterns and 1317 entries contain class based information. The number of rigid gap patterns is 1030. The average PPV is 95.92% and the average Sensitivity is 90.16%. The overall average gap length of the motifs is 1.93 and the standard deviation is 1.52. For the universe of protein sequences we use the Swiss-Prot database [12] (release 49.0). This database contains more than 8 millions of amino-acids for a total of 207132 non-redundant protein sequences.

A. Correlation Analysis

As a first experiment, we evaluated the correlation degree between the measures. For all the 1330 prosite patterns the score of the different measures was calculated. This results in a vector of values per measure. Next, an all-against-all vector comparison was made and the respective correlation calculated. The correlation matrix is plotted in Figure 1. Dark areas indicate a high correlation.

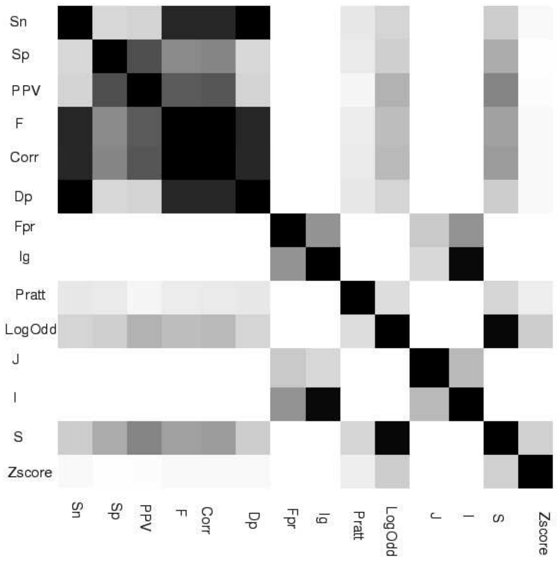

Fig. 1. Plot of the Correlation Matrix.

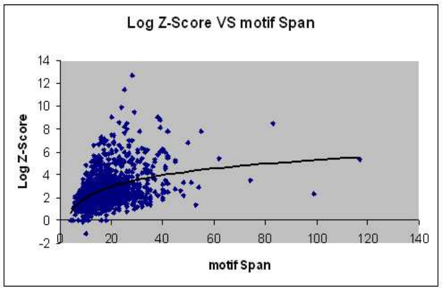

Fig. 2. Plot of the log of Z-Score versus the motif span.

From Figure 1, one can conclude that the class-based measures, with the exception of Fpr, show a high inter correlation. The scatter plots for these measures (Figure 3) shows that these correlations tend to be positive.

Biological sequence databases are often characterized as being highly class imbalanced, where the majority of the cases are negative. In those cases, measures that make use of negative information, as Fpr or specificity, are not suitable. The analysis of the Fpr scores shows that all motifs score closer to zero. This negative rank is of no use in this context, since no discrimination among the patterns can be obtained. In the same way, specificity will always show high scores due to large T_N values.

Further analysis also shows interesting positive correlations among I and IG, logOdd and S, S with PPV and Corr. From the scatter plot in Figure 4 a negative correlation among J with LogOdd and J with S is found.

From manual inspection of the data, we verified that the Z-Score measure shows an almost linear correlation with the motif span. Figure 2 shows this relation. Since Z-Score achieves high values, the logarithm is used instead of the actual score values.

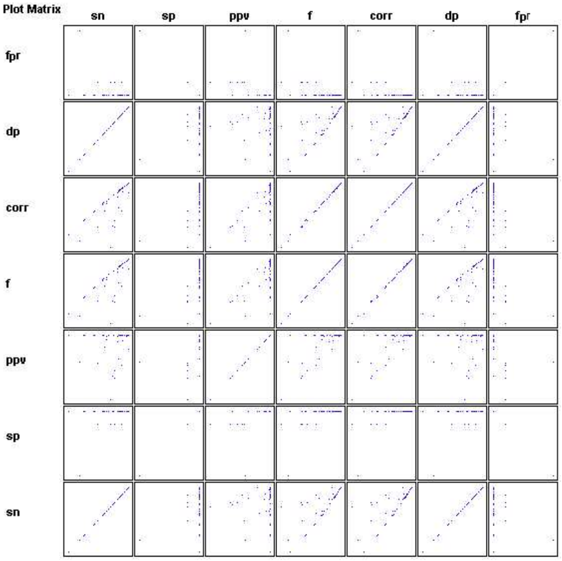

Fig. 3. Scatter Plot of the first 7 measures.

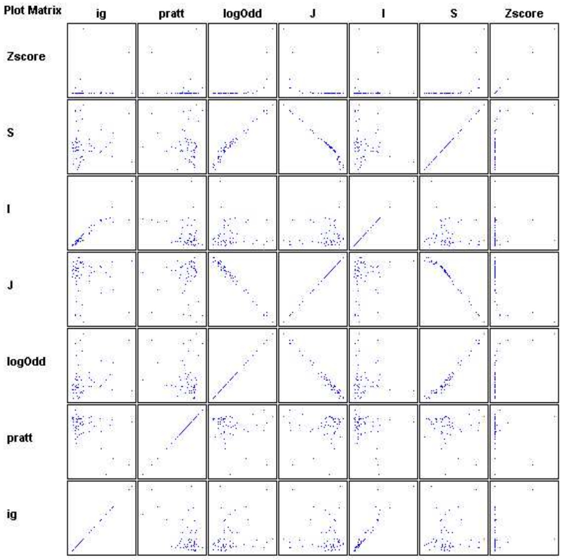

Fig. 4. Scatter Plot of the last 7 measures.

B. Principal Component Analysis

We make use of the Principal Component Analysis - PCA [31], [33] technique to summarize and discover patterns of inter-correlations among the studied measures. This method describes the variation of a set of correlated variables in terms of a set of uncorrelated combinations, called principal components. These components, which express combinations of the original variables, allow a dimensionality reduction while maintain as much as possible of the original variation.

After the application of the PCA method we obtain 14 components, where 4 have an initial eigenvalue greater than 1. Figure 5 shows the scree plot for the 14 components. The first four components show the highest percentage of variance and account for a cumulative variance of 89.1%. We have applied a rotation to the component matrix, see Figure 6 for a 3D visualization of the three components, according to varimax method with Kaiser Normalization [33]. Using a threshold value of 0.5, we can see that in component one the measures L, S and Z-score are highly correlated. In component two are four measures: Sn, F, C and Dp. Note that in Figure 6 the two higher points contain an overlap of F and C and Sn and Dp measures. In component three: Sp, PPV, F and C. Finally in component four: IG and I are highly correlated. Component 2 and 3 relates only class-based measures, where F and C measure are present in both components. This is due to the high inter-correlation of these two measures and the high correlation of the other measures of this class. The two remaining components more surprisingly interesting. Component four relates IG and I measures which are two completely different measures. IG does not make use of any class information and I is essentially class based. Component one relates measure L and Zscore, where some relationship can be found since both provide a degree of emergence of the pattern, i.e. how much its appearance deviates from what was expected. These two measures are also correlated with the S measure, which combines positive predictive value with the information content of the pattern. In this case no obvious mathematical relation can be found between S and the L and Z-score measures.

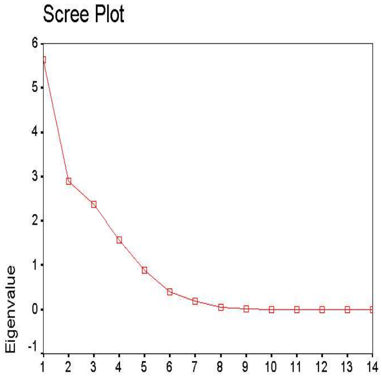

Fig. 5. Scree Plot of the 14 components.

C. Variation Analysis

Since significance measures are used as discrimination tools, an important property of a measure is its variability. Table III shows the average, standard deviation and the coefficient of variation [31] for each measure. From this table we can see that Z-score shows an extremely large variation, due to the existence of extremely large values. The IG and Pratt measure also show a considerable coefficient of variation. The class-based measures shows small values of variation, but one should remember that for those cases the domain range is very limited.

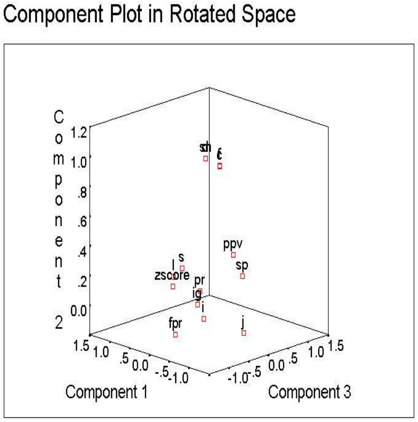

Fig. 6. Plot of the Rotated Components.

Measure	Avg	Std	$\frac{Std}{Avg}$
Sn	0.910	0.122	0.134
Sp	1.000	0.000	0.000
PPV	0.968	0.091	0.094
F	0.931	0.099	0.106
Corr	0.935	0.091	0.097
Dp	0.919	0.122	0.132
Fpr	0.000	0.000	0.000
IG	552.031	755.787	13.733
PR	-3.615	45.055	12.463
L	3.736	3.002	0.817
J	-8.888	3.119	0.359
I	0.005	0.007	1.400
S	7.467	2.612	0.349
Zscore	3M	124M	41.3M

TABLE III
AVERAGE, STANDARD DEVIATION AND COEFFICIENT OF VARIATION OF THE MEASURES.

V. DISCUSSION

In this study a general purpose evaluation of the significance measures has been made. Since different measures have different properties, the best measures or set of measures should be selected according to the problem being tackled. At the moment of selecting those measures caution has to be taken in order to avoid biased results.

When performing a global comparison, we look for the fulfillment of two desirable properties. First, measures should show low correlation with other measures, since two highly correlated measures are redundant. Second, because the evaluated motifs have completely different characteristics in terms of amino-acid composition, number and length of gaps, number of don't cares symbols, motif length and so on, they should provide considerable different scores among the evaluated patterns. Therefore a significant variability should be verified.

In general we can say that class-based measures are significantly correlated. The measures Sn, Sp, PPV, Dp overemphasize some aspects of the pattern quality and should be combined since they do not work well as single evaluators. If class information is required a combination of the Sn and the PPV measures can be used.

They tend to have a relative small correlation and therefore cover different quality aspects of the motif. In those cases, where only one score value can be retrieved, the Correlation measure is recommended. Although this last measures has a high correlation with F and D and approximately the same variability it provides a more balanced evaluation since it makes uses of the four class-based parameters.

The remaining measures evaluate different aspects of the pattern quality and therefore should be chosen according to the target application. Considering the two above stated properties we can say that Z-score, and Pratt are the measures that best fulfill the criteria expressed by these properties.

The present study was intended to provide guidelines for choosing the best set of significance measures. The ideal benchmark would consist in analyzing and comparing how the score provided by the measures agree with the functional sites of the protein sequence. As a future work, we plan to study how these measures can be used for the discovery of such sites. We would also like to study the impact of such measures for specific tasks like classification and clustering.

Acknowledgment

The authors acknowledge the support of the "Fundação para a Ciência e Tecnologia", Portugal. Pedro Gabriel Ferreira is supported by the Fellowship (SFRH/BD/13462/2003). Paulo Azevedo is supported by FEDER and Programa de Financiamento Plurianual de Unidades de I & D (FCT).

References

[1] E. Ukkonen, J. Vilo, A. Brazma, and I. Jonassen. Discovering patterns and subfamilies in biosequences. In *Proceedings of Int'l Conference on Intelligent Systems for Molecular Biology, ISMB 1996*.

[2] N. M. Abramson. *Information Theory and Coding*. McGraw-Hill, New York, 1963.

[3] A. Apostolico, M. Comin, and L. Parida. Conservative extraction of over-represented extensible motifs. *Bioinformatics*, 21(1):i9–i18, 2005.

[4] N. Hulo, A. Bairoch, V. Bulliard, L. Cerrutti, E. De Castro, P. Langendijk-Genevaux, et al. The Prosite Database. *Nucleic Acids Research*, 34(Database Issue):D227-D230, 2006.

[5] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bionformatics*, 16(5):412–242, February 2000.

[6] A. Ben-Hur and D. Brutlag. Remote homology detection:a motif based approach. *Bioinformatics*, 19(1):26–33, 2003.

[7] A. Ben-Hur and D. Brutlag. Sequence motifs: highly predictive features of protein function. In *Proceedings of Workshop on Feature Selection, Neural Information Processing Systems*, December 2003.

[8] K. Blekas, D. Fotiadis, and A. Likas. Motif-based protein sequence classification using neural networks. *Journal of Computational Biology*, 12(1):64–82, February 2005.

[9] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. Technical Report 113, Universit of Bergen, Dep. of Informatics, Bergen, Norway, December 1995.

[10] J. Buhler and M. Tompa. Finding motifs using random projections. In *Proceedings of 9th Int'l Conference On Intelligents Systems for Molecular Biology*, Montreal, Canada, April, 22-25, 2001.

[11] E. Eskin, W. Grundy, and Y. Singer. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Journal of Computational Biology*,10(2):187–214, 2003.

[12] Swiss-Prot Protein knowledgebase Expasy. http://www.expasy.org/sprot/.

[13] E. Gasteiger, A. Gattiker, C.Hoogland, I.Ivanyi, R. Appel and A. Bairoch ExPASy: the proteomics serverfor in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13):3784-3788, 2003.

[14] Pedro Ferreira and Paulo Azevedo. Protein sequence classification through relevant sequence mining and bayes classifiers. In *Proceedings of 12th Portuguese Conference on Artificial Intelligence*, Covilhã, December 2005.

[15] G. Bejerano and G. Yona Modeling protein families using probabilistic suffix trees. In ACM press, editor, *In the proceedings of 3rd International Conference on Research in Computational Molecular Biology*, pages 15–24, 1999.

[16] S. S. Henikoff and J. G. Henikoff. Protein family databases. *Encyclopedia of Life Sciences*, 2001.

[17] S.T. Jensen, L. Shen, and J.S. Liu. Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics*, 21(20):3832–9, October 2005.

[18] I. Jonassen, J. Collins and D. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4(8):1587–1595, 1995.

[19] Anders Krogh. *Computational Methods in Molecular Biology*, volume 1, Chapter 4 - An Introduction to Hidden Markov Models for Biological Sequences, pages 45–63. Elsevier, 1998.

[20] A. Krogh, M. Brown, I. Saira Mian, K. Sjolander and D. Haussler Hidden markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, (235):1501–1531, 1994.

[21] Stefano Lonardi. Tutorial - pattern discover in biosequences. *10th Int'l Conference on Intelligent Systems for Molecular Biology, (Edmonton, Canada)*, August 3-7 2002.

[22] C. Nevill-Manning, K. Sethi, T. Wu, and D. Brutlag. Enumerating and ranking discrete motifs. In *Proceedings of the 5th Int'l Conference on Intelligent Systems for Molecular Biology*, pages 202–209. AAAI Press, 1997.

[23] P. Pevner and S. Sze. Combinatorial approaches to finding subtle signals in dna sequences. In *Proceedings of 8th Int'l Conference On Intelligents Systems for Molecular Biology*, California, USA, 19-23 August 2000.

[24] Marie-France Sagot. On motifs in biological sequences. citeseer.ist.psu.edu/473028.html.

[25] P. Smyth and R.M. Goodman. *Rule Induction Using Information Theory*. MIT press, 1990.

[26] G. Stolovitzky and A. Califano. Statistical significance of patterns in biosequences. Technical report, IBM Computational Biology Center, October 1978.

[27] Ian Witten and Eibe Frank. *Data mining*. Morgan Kaufmann Publishers, 2001.

[28] Thomas D. Wu and Douglas L. Brutlag. Identification of protein motifs using conserved amino acid properties and partitioning techniques. In *3rd Int'l Conference on Intelligent Systems for Molecular Biology*, pages 402–410, 1995.

[29] J. Yang, P. Yu, and W. Wang. Mining surprising periodic patterns. In *Proceedings 7th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, August 2001.

[30] Jiong Yang and Wei Wang. Cluseq: Efficient and effective sequence clustering. In *Proceedings of the 19th Int'l Conference on Data Engineering*, Bangalore, India, March 2003. IEEE Computer Society.

[31] J. H. Zar. *Biostatistical Analysis 3rd Edition*. Prentice Hall, 1999.

[32] E. Koonin and M. Galperin. *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, 2003.

[33] Andy Field. *Discovering Statistics Using SPSS, 2nd Edition*. Sage Publications Ltd, 2005.

[34] B.B. Matthews Comparison of predicted and observed secondary structure of t4 lysozyme. *Biochimica et Biophysica Acta*, 405:442-451, 1975.

[35] G. van den Eijkel. *Intelligent Data Analysis, in M. Berthold and D. Hand (Eds.), 2nd Edition*. Appendix B: Information-Theoretic Tree and Rule Induction, pages 465-463. Springer, 2003.

[36] M. Bramer. Using J-pruning to reduce overfitting in classification trees. *Knowledge-Based Systems*, 15(5-6):301-308, 2002.

Symbol	Measure	Formula	Range	Type
Sn	Sensitivity	$Sn(M) = \frac{T_P}{T_P+F_N}$	[0,1]	1
Sp	Specificity	$Sp(M) = \frac{T_N}{T_N+F_P}$	[0,1]	1
PPV	Positive Predicted Value	$PPV(M) = \frac{T_P}{T_P+F_P}$	[0,1]	1
Fpr	False Positive Rate	$Fpr(M) = \frac{F_P}{F_P+T_N}$	[-1,1]	1
F	F-Measure	$F(M) = \frac{2 \times Sensitivity \times PPV}{Sensitivity+PPV} = \frac{2 \times T_P}{2 \times T_P+F_N+F_P}$	[0,1]	1
C	Correlation	$C(M) = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P+F_N)(T_P+F_P)(T_N+F_P)(T_N+F_N)}}$	[-1,1]	1
Dp	Discrimination Power	$Dp(M) = \frac{T_P}{\lvert C \rvert} - \frac{F_P}{\lvert \bar{C} \rvert}$	[-1,1]	1
IG	Information Gain	$IG(M) = I(M) \times [Support(M) - 1]$ where $I(M) = -log_{\lvert\Sigma\rvert} P(M)$	$[0, +\infty]$	2
PR	Pratt Measure	$PR(M) = \sum_i^n I'(A_i) - c \cdot \sum_{k=1}^{n-1} (q_k - p_k)$ where $I'(A_i) = -\sum_{a_i \in A_i} (P(a_i) \times log(P(a_i))) + \sum_{a_i \in A_i} (\frac{P(a_i)}{P(A_i)} \times log(\frac{P(a_i)}{P(A_i)}))$ and $P(A_i) = \sum_{a_i \in A_i} p(a_i)$	$[-\infty, +\infty]$	2
L	Log-Odds	L(M) = $log(\frac{\frac{Support(M)}{TotalNumPatterns}}{P(M)})$	$[-\infty, +\infty]$	2
Zscore	Z-Score	$Zscore(M) = \frac{Support(M)-E(M)}{N(M)}$ where $E(M) = N_{resid} \times P(M)$ and $N(M) = \sqrt{N_{resid} \times P(M) \times (1 - P(M))}$	$[-\infty, +\infty]$	2
J	J-Measure	$J(C; M) = P(M) \times j(C; M)$ where $j(C; M) = P(C\|M) \times log_2 \frac{P(C\|M)}{P(C)} + (1 - P(C\|M)) \times log_2 \frac{(1-P(C\|M))}{(1-P(C))}$	$[0, 1]$	3
I	Mutual Information	$I(Q; M) = H(Q) - H(Q\|M)$ where $H(Q) = -\sum_{q \in \{C, \bar{C}\}} P(q) \times log_2 P(q)$ and $H(Q\|M) = -P(M) \times \sum_{q \in \{C, \bar{C}\}} P(q\|M) \times log_2 P(q\|M)$	$[0, 1]$	3
S	Surprise Measure	$S(M) = I(M) \times P(C\|M) = I(M) \times \frac{Support(M \cup C)}{Support(M)} = I(M) \times \frac{T_P}{T_P+F_P}$	$[0, +\infty]$	3

TABLE I

LIST OF THE MOTIF SIGNIFICANCE MEASURES.