Bioinformation
www.bioinformation.net

open access
Hypothesis

Prediction of functional engrailed homology-1 protein motif from sequence

Danielle S. Dalafave

Department of Physics, The College of New Jersey, Ewing, New Jersey, USA; Danielle S. Dalafave – Email: dalafave@tcnj.edu; *Corresponding author

Received June 26, 2009; Revised August 10, 2009; Accepted September 11, 2009; Published December 02, 2009

Abstract:
Prediction of functional peptide motifs from sequences is an important problem in bioinformatics. Experimental discovery of functional sequences is laborious. Searches for specific motifs within the increasing number of proteins available in public databases often involve extensive computer calculations. Short peptide motifs are especially hard to identify via currently available methods. Presented here is a simple and effective procedure to identify a short functional motif. The procedure is based on devising a scoring function using sequence properties. The procedure was applied on short engrailed homology-1 (eh1)-like motif. Eh1-like motif provides repressive functions by binding to the WD domain of the Gro/TLE transcriptional corepressors. Interactions of known eh1-like variants and the WD domain were modeled and studied. Sequence features crucial for the interactions, and thus the motif's functionality, were identified. A scoring function was devised based on the observed sequence features. The ability of the scoring function to discriminate between functional and nonfunctional sequences was tested on known eh1-like sequences, random sequences, and eh1-like sequences predicted by others using various bioinformatics tools. The scoring function expressed well a general relationship between sequences and their functionalities. It gave about 20% false positive findings. However, the scoring function reliably identified sequences that were not functional eh1-like motif. The procedure presented here may prove useful for predicting functional sequences of other short motifs. Given the importance of transcriptional regulation, this study on identification and evaluation of functional eh1-like sequences should facilitate further research on their transcriptional roles.

Background:
Development of dependable computational methods to identify functional peptide motifs is an important quest in bioinformatics. Experimental discovery of functional peptide sequences is often arduous. High-throughput methods give a large number of false positive and false negative results. Extensive computer calculations are often needed to search for specific motifs within protein databases. Identification of short motifs using currently available methods is especially difficult. Multiple sequence alignments are tools used in several motif-searching programs. One of them is a powerful program Clustal W **[1]**. However, aligning sequences by Clustal W requires significant computational resources. It can take hours of sequential computing to align a few hundred sequences.

BLAST **[2]** is widely used for matching a peptide sequence with database sequences and reporting results above a specified threshold. Stressing speed over sensitivity, BLAST looks only for significant patterns within the sequences. Another program, MEME **[3]**, is used to search for new motifs in user-supplied protein sequences. Various online databases store a few hundred currently known peptide motifs. Examples are ELM **[4]** and SCANSITE **[5]**. To decrease the number of false-positive results, the ELM filter hides protein sections in which motifs are unlikely to be present. SCANSITE employs position-specific scoring matrices to evaluate functionality of putative motifs. However, the matrices can be created only for motifs for which many sequences have been experimentally confirmed as functional. When searching for motifs less than 10 residues long, these techniques usually give a large number of false-positive results.

Presented here is a procedure to identify short functional sequences using the sequences' amino acid residues. The method is simple, yet effective, and requires minimal computer power and programming skills. The method can help reduce a pool of putative motif candidates that warrant further experimental or theoretical investigations. It also provides an additional tool to validate functionality of sequences found by other bioinformatics methods. In this work, engrailed homology-1 (eh1)-like motif was investigated. The motif's transcriptional role, its short length, and a relatively small number of sequences that have been experimentally confirmed as functional, were some of the main reasons the motif was chosen for the study.

Transcription factors help control transcription of genetic information from DNA to RNA. Gro/TLE proteins are corepressors for various transcription factors. Many of these factors interact with Gro/TLE through eh1-like motif **[6]**. During the interaction, eh1-like motif binds to the Gro/TLE via its WD domain. The WD domain is a highly conserved C-terminal region of Gro/TLE. As demonstrated by X-ray studies **[6]**, the WD domain forms a β-propeller structure. Specific transcription factors are recognized by specific propeller's binding sites. Eh1-like motif conforms to the consensus pattern FΣBXXBBX **[7]**. Here, F is phenylalanine; Σ = S or T; B = branched hydrophobic residue (I, L, or V); and X = nonpolar residue (A, D, E, F, G, H, I, K, L, M, P, R, V, or W). Substitutions with 'non-consensus residues' are allowed up to a point and can lead to functional sequences **[8, 9]**.

Literature searches identified fifty experimentally-confirmed eh1-like sequences **[8-12]**. Copley **[13]** predicted 109 eh1-like sequences using BLAST database searches. Yaklichkin et al. **[7]** searched for novel motifs among 458 metazoan proteins using Clustal W and MEME programs. They predicted 95 new eh1-like sequences. All of the above sequences were used to train or test the prediction method described here. A task addressed in this paper may be summarized in the following way. Given a set of peptide sequences, develop a method to determine whether each sequence is a functional or nonfunctional eh1-like motif. A simple, yet effective procedure will be described, which identifies functional eh1-like motif variants from their sequences with high reliability.

Methodology:
Dataset:
About 1000 sequences were used to train and test the scoring function. The training set consisted of twenty experimentally known eh1-like sequences **[9-12]** and forty random eight-amino-acid long sequences. The random sequences were checked against protein databases to make sure they were not functional. Four sets of sequences were used to test the scoring function. The first set included 30 experimentally known eh1-like variants **[9-12]**, none of which was in the training set. The second set contained about 800 random eight-amino-acid long sequences. The third had 109 eh1-like sequences predicted by Copley **[13]** and the fourth, 95 eh1-like sequences predicted by Yaklichkin et al. **[7]**.

Procedure employed:

The scoring function:

The objective was to devise a scoring function that would discriminate between functional and nonfunctional sequences using sequence features. The scoring function of the form $F = \Sigma w_i \bullet f_i$ was used. Here, w_i were weighing coefficients and f_i's quantified relevant sequence properties. The procedure for determining wi's and f_i's is described below. A starting assumption was that residues at positions in agreement with the consensus sequence had a positive influence on the sequence's functionality. These will be referred to as 'positive-influence' (PI) residues. The training sequences were scored using f_i = 1 when a PI residue was present and f_i = 0 when it was absent from a given position.

Determination of sequence properties:

The next step was to determine sequence properties relevant to motif's interactions with the WD domain, and thus to motif's functionality. Experimentally determined three-dimensional (3D) models of interactions between the WD domain and the 20 functional training sequences were not available in protein databases. Therefore, the models were created and optimized via Deep View program and the Swiss Model **[14]**. Experimentally known 3D structure of an eh1-like sequence bound to the WD domain (PDB ID: 2CE8) was used as a template. The sequence's residues were mutated in Deep View to obtain, in turn, models of interactions for each of the 20 sequences. The models were optimized in Deep View and submitted to the Swiss Model. Final models were built and returned to Deep View. The models of functional sequences helped identify both essential and nonessential binding sites for the WD-eh1 interactions. Residues at each position were systematically mutated in each model to see whether the interactions with the WD domain were eliminated. Residues at particular positions that abolished WD-eh1 interactions will be referred to as 'negative-influence' (NI) residues.

Tuning of the scoring function:

The scoring function was then amended to include information on sequence properties found to be crucial for the functionality. The training sequences were scored using f_i = -1 when an NI residue was present. When an NI was absent, f_i was 1 for that position, since the absence of an NI residue increased the sequence's chance of being functional. To assess the influence of each term on the scores, the weights were systematically changed for each position. If the change had a similar effect on both functional and nonfunctional sequences, and the modeling indicated the position to be nonessential, that term was removed from the scoring function.

Evaluation of the scoring function:

The scoring function's performance was assessed using the four sets of test sequences. The performance was also compared to that of an online server EasyPred **[15]**.

Discussion:

Computer modeling indicated that WD residues D617, R534, and E550 interacted, respectively, with residues 1, 2, and 3 in most functional eh1-like sequences. Most sequence positions could be occupied by a limited number of residues with similar hydrophobicity, charge, etc., and still preserve functionality. Findings on positional distributions of the residues were summarized in the Table below. Residues not listed did not significantly contribute and were omitted from the scoring function.

As shown in the Table, the WD-binding positions 1, 2, and 3 did not accept charged residues D, R, E, and K. Also, residue H was forbidden in positions 1, 3, 6, and 7. Most residues, except I, L, and V, were not allowed in the 6th position. Position 8, which did not interact with the WD domain, could accept any amino acid, as could position 4. Flexibility of position 4 may be related to its residue serving to ensure proper spacing between key amino acids.

The clearest distinction between functional and nonfunctional sequences was achieved with this scoring function:

F = 1•(PI or NI in position 1) + 1•(PI or NI in position 2) + 0.8•(PI in positions 3, 4, or 6) + 0.7•(PI in position 5) + 0.5•(PI in positions 7 or 8) + 0.8•(NI in position 5) + 0.1•(NI in positions 3, 6, or 7)

where PI = {1 or 0} corresponded to 'positive influence' terms and NI = {-1 or 1} to 'negative'. All weights were normalized to position 1 coefficient. There were 14 term in the formula. Since each position was occupied by only one residue at a time, only 8 addends contributed to the final score.

The scoring function agreed with findings from the computer modeling. Positions 1 and 2 were important for the scores, as can be seen from their weights. If non-consensus residues were in these positions, the sequence was likely nonfunctional. Also important were PI residues in positions 3 and 6, and to a lesser extent in position 5. Positions 4 and 8 did not have any NI terms, which is consistent with the Table.

The scoring function captured a general correspondence between the likelihood that the sequence was functional and sequence's characteristics. The scoring function gave no false negative results. That is, 100% of the experimental eh1-like sequences **[8-12]** received 'functional' scores. About 91% of sequences predicted by Copley **[13]** and about 95% of those predicted by Yaklichkin et al. **[7]** were identified as functional by the scoring function. Most random sequences received nonfunctional scores. The scoring function gave about 20% false positive results, identifying about 20% of random sequences as functional. When checked against protein databases most of these sequences were not found, indicating that they were not known to be functional. Several false positive sequences that were found in the databases were present inside proteins' DNA-binding domains. Due to steric constraints, a functional eh1-like motif is expected to be outside the DNA-binding domain. Indeed, previously identified functional sequences **[8-12]** were all found outside this domain. It was concluded that these false positives were not functional sequences.

The model's performance was compared to that of EasyPred server **[15]**. EasyPred was trained with the same set of 20 functional sequences. It predicted all sequences identified by Copley **[13]** and Yaklichkin et al. **[7]** as functional. This is not surprising, since they used similar prediction methods and online resources. Although EasyPred correctly predicted functionality of all experimental sequences, it gave 35% false positives among random sequences. This is compared to 20% obtained with the method presented here. About 93% of scoring terms used in this method agreed with corresponding terms in EasyPred scoring matrix. In addition, this model is simpler since it uses 14 weight coefficients, compared to 160 in EasyPred matrix.

A ROC plot was used to further evaluate the model's performance (**Figure 1**). In the plot, a sensitivity increase corresponds to a specificity decrease. The sensitivity is the fraction of correctly scored known positives (true positive rate or TPR). The specificity is 1- FPR, where FPR (false positive rate) is the fraction of incorrectly scored known negatives.

The closer the ROC curve comes to the y-axis and to the top of the plot, the more accurate the model is. An ideal model would give TPR = 1 and FPR = 0. The optimal threshold, which maximizes both the model's sensitivity and specificity, corresponds to the beginning of the plateau of the curve. For this model, the optimal threshold was about FPR = 0.2.

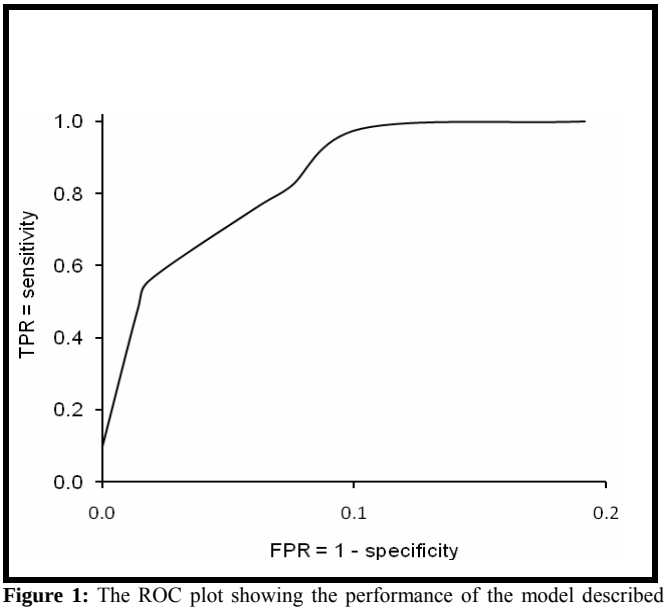

Figure 1: The ROC plot showing the performance of the model described in the paper. The curve represents the portion of correctly scored known positives (true positive rate, TPR) vs. the portion of incorrectly scored known negatives (false positive rate, FPR).

Conclusions:

An important problem of determining functionality of peptide sequences from their amino acids is addressed here. Interactions of functional eh1-like variants with the WD domain were modeled. A scoring function was devised and tested on known, predicted, and randomly generated sequences. The scoring function was able to discriminate between functional and nonfunctional sequences. Although it gave about 20% false positive findings, the function reliably identified nonfunctional sequences. The model was more successful than an online server in reducing the number of false positives. At least 91% of sequences previously predicted to be functional by other bioinformatics methods were identified as functional using this model.

The strength of the model is its simplicity and effectiveness. One of the main advantages is that good results can be obtained using modest computer power and minimal programming skills. The model was developed using a relatively small number of functional training sequences. Also, the 3D models of sequences were obtained computationally, without experimental confirmation. These may have lead to missing some NI or PI residues and to inaccuracies when determining the weights. However, the success of the model suggests that it can be valuable for other short motifs with small number of experimentally available sequences and structures.

Further studies of selective protein recognition building on this method are planned. Given the importance of transcriptional regulation, this study on functionality of eh1-like sequences should facilitate further research on their transcriptional roles.

References:

[1] JD Thompson *et al., Nucleic Acids Res.* **22**: 4673 (1994)
[2] SF Altschul *et al., Journal of Molecular Biology* **215**: 403 (1990)
[3] TL Bailey, C Elkan, *Mach Learn.* **21**: 51 (1995)
[4] P Puntervoll *et al., Nucleic Acids Res.* 31: 3625 (2003)
[5] J Obenauer *et al., Nucleic Acids Res.* **31**: 3635 (2003)
[6] BH Jennings *et al., Mol Cell.* **22**: 645 (2006)
[7] S Yaklichkin *et al., BMC Genomics* **8**: 201 (2007)
[8] RE Goldstein *et al., Mol Cell Biol.* **25**: 10711 (2005)
[9] ST Smith, JB Jaynes, *Development* **122**: 3141 (1996)
[10] D Eberhard *et al., EMBO J.* **19**: 2292 (2000)
[11] CC Zhu *et al., Development* **129**: 2835 (2002)
[12] HF Farin *et al., J Biol Chem.* **282**: 25748 (2007)
[13] RR Copley, *BMC Genomics* **6**: 169 (2005)
[14] http://spdbv.vital-it.ch/; http://swissmodel.expasy.org//SWISS-MODEL.html
[15] http://www.cbs.dtu.dk/biotools/EasyPred/

Edited by P. Kangueane
Citation: Dalafave, Bioinformation 4(6): 229-232 (2009)

Supplementary material:

Table 1: Positional distribution of residues within eh1-like motif. Residues that did not significantly contribute to the scoring function are not listed.

Position within the sequence	PI residues	NI residues
1	F	E, D, G, H, K, P, R
2	S, T	D, E, K, R
3	I, L, V	D, E, G, H, K, R
4	D, E, K	
5	A, D, E	T
6	I, L, V	A, C, D, E, G, H, K, M, N, P, Q, R, S, T, Y
7	L, V	N, H, Q, R, S, T
8	D, R	