Prediction of nuclear export signals using weighted regular expressions (Wregex)

Gorka Prieto 1,* Asier Fullaondo 2 and Jose A. Rodriguez 2

1 Department of Communications Engineering, University of the Basque Country (UPV/EHU), Alda. Urquijo s/n Bilbao, 48013, Spain
2 Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n Leioa, 48940, Spain

ABSTRACT

Motivation: Leucine-rich nuclear export signals (NESs) are short amino acid motifs that mediate binding of cargo proteins to the nuclear export receptor CRM1, and thus contribute to regulate the localization and function of many cellular proteins. Computational prediction of NES motifs is of great interest, but remains a significant challenge.

Results: We have developed a novel approach for amino acid motif searching that can be used for NES prediction. This approach, termed Wregex (weighted regular expression), combines regular expressions with a Position-Specific Scoring Matrix (PSSM), and has been implemented in a web-based, freely available, software tool. By making use of a PSSM, Wregex provides a score to prioritize candidates for experimental testing. Key features of Wregex include its flexibility, which makes it useful for searching other types of protein motifs, and its fast execution time, which makes it suitable for large-scale analysis. In comparative tests with previously available prediction tools, Wregex is shown to offer a good rate of true positive prediction for large-scale analysis. In comparative tests with previously available protein motifs, and its fast execution time, which makes it suitable for searching other types of NESs. One of the best-studied types of nuclear export signals is the leucine-rich NES that mediates binding of cargo proteins to the nuclear export receptor CRM1 (Hutten and Kehlenbach, 2007).

A leucine-rich NES typically consists of a short stretch of amino acids containing several hydrophobic residues (not necessarily leucine) with a characteristic pattern. Initial studies of a limited number of NESs (Boger et al., 1996; La Cour et al., 2004) led to propose a consensus NES pattern $\Phi^1 - (X)_{2-3} - \Phi^2 - (X)_{2-3} - \Phi^3 - (X) - \Phi^4$, with four conserved hydrophobic residues (represented by $\Phi^{1-4}$) separated by a variable number of intervening residues (represented by X). This consensus pattern has been progressively expanded and refined through the analysis of a larger number of experimentally confirmed NESs (Kosugi et al., 2008; Fu et al., 2011; Xu et al., 2012). These studies, mostly based on sequence alignment, suggested that residues at positions other than the conserved $\Phi^{1-4}$ may also contribute to NES activity. For example, acidic residues Glu, Asp were found to be overrepresented in functional NESs (La Cour et al., 2004; Xu et al., 2012). Further insight into NES features has been recently obtained through structural analyses of the CRM1-NES binding interface (Monecke et al., 2009; Gütter et al., 2010). The crystal structures of CRM1-NES complexes showed that NESs fit into a rigid hydrophobic groove on the surface of the receptor, and that a fifth hydrophobic residue ($\Phi^5$) in the NES, amino-terminal to $\Phi^1$, may increase the affinity of the NES/CRM1 interaction (Monecke et al., 2009; Dong et al., 2009; Gütter et al., 2010). These structural analyses also revealed that acidic residues in certain positions may contribute to NES affinity by interacting with a basic surface flanking CRM1 hydrophobic groove (Dong et al., 2009), thus providing a biological explanation to the previously noted overrepresentation of acidic amino acids in NESs. Altogether, these previous studies have unveiled the remarkable complexity of the CRM1-dependent NES.

In addition to the complex nature of the signal, the identification of physiologically relevant NESs is further hampered by the fact that amino acid motifs that conform to the NES pattern may be found in hydrophobic protein cores, where they would be unavailable for CRM1 binding. A valid strategy for NES identification may begin by mapping amino acid segments with CRM1-dependent export activity (active NES motifs) in the protein of interest. If the three-dimensional structure of the protein is known, the solvent

*To whom correspondence should be addressed

© The Author (2014). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
accessibility of these motifs may then be evaluated and finally, their physiological relevance can be validated by mutagenesis. In the first step of this strategy, the use of efficient computer-based NES prediction tools may play an important role. Several NES-prediction tools have been developed to date. These include programs that apply regular expressions derived from the NES consensus pattern, such as ELM (Gould et al., 2010) as well as machine learning-based programs, such as NetNES (La Cour et al., 2004) and NESSential (Fu et al., 2011).

We have recently compared the performance of ELM and NetNES using these programs to predict candidate NES motifs in human deubiquitinasases (DUBs) (García-Santisteban et al., 2012). The more recent NESSential program was not available at the time this comparison was carried out. Predicted candidates were subsequently tested using an in vivo nuclear export assay (Henderson and Eleftheriou, 2000). In our hands, the ability of ELM to identify amino acid motifs with nuclear export activity was superior to that of NetNES, although the estimated positive predictive value (26.0% for NetNES and 38.0% for ELM) was relatively low for both programs (García-Santisteban et al., 2012).

We have devised and tested a novel approach, termed weighted regular expressions (Wregex), that can be applied to the prediction of functional amino acid motifs, including NESs. Wregex combines the use of a regular expression with a Position-Specific Scoring Matrix (PSSM). In the case of NES prediction, the PSSM takes into account the potential contribution of each NES residue (not only those at the conserved hydrophobic positions) to NES activity.

In this study, we submit Wregex to a series of comparisons with NESSential and ELM, and progressively optimize its regular expression and training PSSM. The features of Wregex in comparison to these existing tools, as well as the potential application of Wregex to the identification of other functional motifs in proteins besides NESs are discussed. We have implemented Wregex as a web interface accessible to any researcher.

2 METHODS

2.1 Wregex in the context of protein and motif searching tools

Figure 1 summarizes the approaches used by different protein and motif searching tools. Tools based on sequence similarity like BLAST (Altschul et al., 1997) and FASTA (Pearson, 1990) use a substitution matrix for scoring alignments and allow specifying a sequence gap penalty. These tools are more indicated for comparing the similarity between proteins rather than for motif searching. PSI-BLAST (Altschul et al., 1997) uses an initial protein–protein BLAST to derive a PSSM, which is then used for further database search in an iterative manner updating the PSSM in each iteration. Rather than detecting specific protein motifs, the aim of PSI-BLAST is to detect distant relationships between proteins. PHI-BLAST (Altschul et al., 1997) uses both a protein sequence and a PROSITE (Hulo et al., 2006) regular expression as input. The regular expression is used to search the database for proteins that are subsequently aligned with the input protein sequence using BLAST to obtain a score, which is computed using a substitution matrix like BLOSUM-62 instead of a PSSM.

Another motif searching approach consists on machine learning using a training set of sequences. The most commonly used algorithms in this approach are hidden Markov model (HMM) and support vector machine (SVM). The NES prediction tool NetNES (La Cour et al., 2004) belongs to this category, and uses a combination of both neural networks and HMM.

A more recent NES prediction program, NESSential (Fu et al., 2011), uses SVM models trained with sequence derived meta-features, and a pre-filter consisting on the pattern $\Phi - (X)_{2-3} - \Phi - (X) - \Phi$. This tool will be further discussed below.

ScanProsite (De Castro et al., 2006) allows to scan the PROSITE database (Hulo et al., 2006) using either patterns (regular expressions) or profiles (tables of position-specific amino acid weights and gap costs), but it does not allow to combine both a regular expression and a weight matrix. Finally ELM (Gould et al., 2010) is a database of eukaryotic linear motifs described as regular expressions, including the “TRG_NES_CRM1_1” that allows prediction of CRM1-dependent NES motifs.

The novel motif searching tool presented in this paper, Wregex, combines both a regular expression and a PSSM.

2.2 Algorithm implementation

Wregex uses three type of inputs:

1. Regular expression (mandatory). Defines the sequence patterns and spacings allowed. This expression allows the use of “(brackets)” to define the groups of amino acids that will be weighted together later.
2. PSSM (optional). Defines the weight of each possible amino acid for each of the groups marked between “(brackets)” in the regular expression. These weights are interpreted as the base 10 logarithm of the amino acid probability. If no PSSM is provided, score of output candidates will not be available.
3. Fasta database (mandatory). The protein sequence database used for searching protein motifs using the regular expression and the PSSM (if provided).

Figure 2 summarizes how Wregex uses these inputs for searching candidates. Only amino acid sequences matching the regular expression will be subsequently processed. This selection represents a first filter for candidate motifs. Next, each of the candidate sequences is subdivided into smaller amino acid groups according to the regular expression, and a score is computed using the PSSM values for each group position. A threshold (arbitrarily selected by the user) can be applied to these scores, thus introducing a second filter for candidate NES motif selection.

Equation 1 is used for computing the score of a motif match; where $G$ is the number of groups in the regular expression, $j_g$ is the length of the group number $g$, $P_g$ is the column $g$ of the PSSM matrix, $p_{g,1|j_g}$ is the PSSM
Fig. 2. Motif search and score computation in Wregex combining a regular expression with a Position-Specific Scoring Matrix (PSSM).

Fig. 3. Position-Specific Scoring Matrix (PSSM) calculation using experimentally validated NES motifs with an activity score determined using an export assay. Sequences are first aligned using the groups in the regular expression. The weight of a motif (\(W\)) is derived from the assay score (\(S\)). If several amino acid combinations in a motif match the regular expression, the final weight (\(W'\)) of the motif is calculated by dividing \(W\) by the number of combinations. When input motifs lacking an activity score are used for PSSM calculation, the same \(W\) value is ascribed to every motif.

value in column \(g\) for the amino acid in the position \(j\) within the group \(g\), and \(len\) is the length of the sequence matched by the regular expression.

\[
\begin{align*}
\text{score} &= \sum_{g=1}^{G} \sum_{j=1}^{J} p_{a,j(g)} \cdot g - \max(P_g) \\
&= 10^\log_{10}(\text{len}) + 100
\end{align*}
\]  

In this equation the group position-specific PSSM value is used as the score of each amino acid of the match. In order to assign all positions the same importance, the score is normalized by the maximum PSSM value of the corresponding group position by subtracting \(\max(P_g)\). Since we are considering variable length matches, \(\text{score}\) must be normalized by the match length \(len\). Finally the logarithm is translated into linear scale resulting in a score ranging from 0 to 1, and then multiplied by 100 to obtain a score ranging from 0 to 100.

Figure 3 illustrates how the PSSM is calculated in the case of experimentally validated motifs with activity score. These motifs are first aligned using the groups in the regular expression. In case that different group combinations are possible, all of them are considered and their weight (\(w\), derived from the activity score (\(s\)), will be divided by the number of combinations, resulting in the final weight (\(w'\)). Next, for each group \(g\) and amino acid \(a\) a \(p_{a,g}\) value is computed using Equation 2; where \(L\) is the number of training matches and \(w_i'\) is the weight of training match \(i\). In the numerator, the weights \(w_i'\) are only added if the amino acid \(a\) is the same as the one in the position \(j\) of the current group \(g\) for the training match \(l\) (\(a_{l,j(g)}\)).

\[
p_{a,g} = \log_{10} \left( \frac{\sum_{l=1}^{L} \sum_{j=1}^{J} a_{l,j(g)} w_i' + 10^{-5}}{\sum_{l=1}^{L} \sum_{j=1}^{J} w_i'} \right) \quad \forall a, \forall g = 1...G
\]  

When calculating a PSSM using validated motifs without assay score, the same weight \(w\) is ascribed to all sequences. In this case, the \(p_{a,g}\) values of the PSSM in Equation 2 represent the probability that amino acid \(a\) is present in group number \(g\). When different weight values \(w\) are considered (derived from assay score), a similar interpretation can be assumed considering that \(w\) represents the number of occurrences of the corresponding training match. These PSSM values are expressed in logarithmic scale to facilitate further operations, and they are rounded to the nearest integer to account for the statistical sample size. The \(10^{-5}\) term is added to avoid logarithm of 0, thus resulting in a minimum value of \(-5\) for \(p_{a,g}\). This \(10^{-5}\) term does not affect significantly the remaining non-zero values (at least two orders of magnitude above in our tests) because of the integer rounding.
The PSSMs used in this paper are calculated using experimentally validated NES motifs. In this regard, it is important to note that those NES motifs that have been validated using the nuclear export assay developed by Henderson and Eleftheriou are assigned an activity score, (ranging from 1+ to 9+) based on the proportion of transfected cells showing nuclear, nuclear/cytoplasmic, or cytoplasmic localization of the Rev(1.4)-NES-GFP protein (Henderson and Eleftheriou, 2000). This score is not available for NES motifs validated differently.

2.3 Web interface

A web application has been developed using Java 7 and JSF 2.2 running in a servlet 2.5 compliant web container. The training page (Figure 4(a)) allows to define a custom regular expression and upload input motifs as fasta sequences. These motifs can be defined as separate fasta entries, or annotated as position ranges in the fasta header of protein entries using the same format provided by ValidNESs (Fu et al., 2013) fasta download. It is also supported to include a motif weight as an additional annotation in the fasta headers by appending “@weight” to the motif position range. Then, a PSSM is computed as described in the previous subsection using those input motifs matched by the regular expression. If several amino acid combinations in a single input motif match the regular expression, the motif weight is divided by the number of combinations. This process can also be tuned by the user by removing undesired combinations. Finally, the PSSM can be downloaded and used later for custom motif searching.

The search page (Figure 4(b)) allows to search a fasta file for protein motifs. The motif can be selected from a predefined set or defined by the user as a custom regular expression with an optional PSSM file. If the input fasta file has header annotations with motifs positions, additional information is provided stating whether the resulting candidates match the annotated motifs (e.g. laboratory assay results). By default, if several motifs candidates overlap in the protein sequence, only the one with the highest score or the longest sequence is displayed. This option can be deactivated using the grouping checkbox. In any case, the number of combinations/overlaps is displayed, as it may be of help when selecting candidate motifs for testing.

2.4 In vivo nuclear export assay

Candidate NESs were tested using an in vivo nuclear export assay (Henderson and Eleftheriou, 2000), as previously described (Garcia-Santisteban et al., 2012). Briefly, double stranded DNA fragments encoding the candidate NES and flanking residues were cloned into the pRev(1.4)-GFP vector (a gift from Dr. Beric Henderson). HeLa cells, growing in Dulbecco’s modified Eagle’s medium, supplemented with 10% fetal bovine serum, 100 units/ml penicillin and 100 µg/ml streptomycin (all from Invitrogen), were seeded onto sterile glass coverslips in 12-well trays. 24 h later, cells were transfected with the pRev(1.4)-GFP-based plasmids containing the candidate NES sequences using X-tremeGENE 9 transfection reagent (Roche Diagnostics) following manufacturer’s protocol. Using a Zeiss Axioskop fluorescence microscope, the subcellular localization of the fluorescent proteins was determined in at least 200 cells per sample, and the activity of the functional NESs was rated between 1+ and 9+ using the scoring system originally proposed (Henderson and Eleftheriou, 2000). Survivin NES (Engelsma et al., 2007) and the empty pRev(1.4)-GFP were used as positive and negative assay controls respectively. A subset of the candidate NESs were also tested in HEK 293 cells, showing equal activity profiles.

3 RESULTS AND DISCUSSION

A prominent feature of Wregex is its flexibility. On one hand, training can be adjusted by defining a “training regular expression” that matches the input motifs used as training sequences for computing the PSSM. On the other hand, searching can also be adjusted for higher or lower stringency, by modifying the “searching regular expression” and selecting a PSSM. Different settings at the level of training or searching result in different Wregex configurations. Here, we have tested several configurations of Wregex (Wregex A-E, summarized in Table 1), in order to carry out a comparison with the previously available NESsential and ELM tools, and in an attempt to obtain an optimal configuration to be used as the “Recommended” default in the Wregex web interface.

The training regular expression is common to all configurations tested here. We have chosen the low stringency regular expression \((\{2\} \{LIVMFAWY\} \{2,3\} \{LIVMFAWY\} \{2,3\}\{P\} \{2,3\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\{2,3\}\{P\}\{LIVMFAWY\}\) for training, in order to maximize the number of training sequences used in the PSSM. As indicated in Table 1 and further detailed below, the differences between the various Wregex configurations lie in the training dataset used for computing the PSSM or in the searching regular expression applied.

3.1 Comparison Wregex A versus NESsential

NESsential (Fu et al., 2011) is the most recently developed NES prediction tool, and it was not tested in our previous comparison of NES predictors (Garcia-Santisteban et al., 2012). Thus, we decided to carry out a comparison between Wregex and NESsential. In order for the comparison to be fair, we trained.
Table 1. Wregex configurations used for comparison with NESsential and ELM, and for training optimization.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Searching reg.ex.</th>
<th>PSSM training dataset</th>
<th>Comparison</th>
<th>Target set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ELM</td>
<td>NESsential training set</td>
<td>Wregex vs NESsential</td>
<td>DUBs</td>
</tr>
<tr>
<td>B</td>
<td>ELM</td>
<td>ValidNESs + DUBs</td>
<td>Wregex vs ELM (prospective)</td>
<td>ASPs</td>
</tr>
<tr>
<td>C</td>
<td>WRE</td>
<td>DUBs + ASPs not using assay scores</td>
<td>Wregex score vs no score</td>
<td>DUBs + ASPs</td>
</tr>
<tr>
<td>D</td>
<td>WRE</td>
<td>DUBs + ASPs using assay scores</td>
<td>Wregex score vs no score</td>
<td>DUBs + ASPs</td>
</tr>
<tr>
<td>E (recommended)</td>
<td>WRE</td>
<td>ValidNESs + DUBs + ASPs</td>
<td>Validation</td>
<td>DUBs + ASPs</td>
</tr>
</tbody>
</table>

\(^a\)Based on ELM TRG NES_CRM1: ([DEQ], {0,1})([LM], {2,3})([LIVMF], {2,3})([LVMF], {2,3})
\(^b\)Proposed NES/CRM1 regex: ([DEQSI], {0,1})([LIMA], {2,3})([LIVMF], {2,3})([LVMF], {2,3})
\(^c\)Arbitrarily selected proteins (ASPs), mostly related to chromatin modification processes

As shown in Figure 5, Wregex A predicted fewer candidate motifs than NESsential (64 vs 826). In order to select a manageable number of candidate motifs to be experimentally tested, a threshold can be arbitrarily defined. As an example, a threshold value of 50 is displayed in Figure 5. Using this threshold, 42 candidates would be selected using Wregex A and 24 using NESsential. Wregex A candidates would include at least 15 true positive and 15 false positive NES motifs, whereas NESsential candidates would include at least 7 true positive and 5 false positive motifs (Table 2). As shown in Table 2, we have also estimated the number of true positives reported by each program at a threshold value (34 for Wregex A, and 14 for NESsential) that leads to the identification of an equal number of true positive motifs (21, the maximum number of true positives reported by Wregex A). Using these thresholds, NESsential reports a higher number of false positives than Wregex A (29 vs 22) and, importantly, a much higher number of candidates that would be selected for experimental testing (214 vs 62). From this comparison, we conclude that Wregex offers a good compromise of true positives with a smaller number of candidates in relation to NESsential.

Another important difference is the computational resources required by these tools. NESsential took several hours to process the proteins used in Figure 5, while Wregex finished in a few seconds.

![Fig. 5. Comparison Wregex A versus NESsential. Graphs show the score assigned by Wregex A (a) or NESsential (b) to each predicted NES motif versus the activity score experimentally assigned to each motif using a nuclear export assay. For displaying purposes, the 0–1 score range provided by NESsential has been scaled to 0–100. Green circles to the right and the left of the threshold indicate true positives and true negatives, respectively. Red circles to the right and the left of the threshold indicate false positives and false negatives, respectively. Black circles represent non-tested (NT) candidates.](http://bioinformatics.oxfordjournals.org/doi/fig/10.1093/bioinformatics/btv156)
This faster execution time would be important for large scale (e.g. proteome-wide) analyses.

### 3.2 Comparison Wregex B versus ELM

The best results in our previously reported comparison of NES predictors were provided by ELM (Garcia-Santisteban et al., 2012). Wregex can be regarded as an evolution of ELM, the main difference being the training of Wregex with a PSSM, which could help refining the regular expression-based search. In order to compare its performance with that of ELM, we used a second Wregex configuration (Wregex B) that uses the ELM-derived expression described above, and is trained with a PSSM (Supplementary Table S2) computed using the experimentally validated NES motifs from the ValidNESs database (Fu et al., 2013) and our previous DUB study (Garcia-Santisteban et al., 2012). We used ELM and Wregex B to predict candidate NES motifs in a set of 21 arbitrarily selected proteins (ASPs) mostly related to chromatin modification processes (Supplementary Table S6). Both tools use essentially the same regular expression, but Wregex allows further selection of candidates by applying the PSSM and the score filter. Thus, ELM predicted 21 candidates, and Wregex, at a score threshold of 50, predicted 16 candidates. These candidate NES motifs were subsequently tested using a nuclear export assay (two motifs could not be analyzed due to cloning problems), and the results are summarized in Table 3 and shown in detail in Table S7. These results indicate that Wregex B, at a threshold value of 50, identifies the same number of true positive NES motifs as ELM (7), while decreasing the number of false positives from 12 to 8, suggesting that the Wregex approach is a valid strategy for NES prediction.

NES prediction with Wregex can potentially be further improved by computing the PSSM with a larger number of experimentally validated sequences, or by adjusting the stringency of the searching regular expression. In this regard, we found that using a more permissive regular expression led to the prediction of three true positive NES motifs (WRE16, WRE19, WRE26 in Supplementary Table S7) that were subsequently experimentally validated in the export assay. Based on this observation, we propose a new searching regular expression

\[
(^[DQE]{1,2}[LIVMF]{1,2}[DEQ]{1,2}[LIV]{1,2})\]

which allows serine before \(\Phi\) and alanine in \(\Phi\), indicating a new regular expression termed WRE (Supplementary Table S7), which allows serine before \(\Phi\) and alanine in \(\Phi\).

### 3.3 PSSM computation using export assay activity score: Wregex C versus Wregex D

In order to gauge the effect of including the activity score for PSSM computation two new Wregex configurations (Wregex C and Wregex D) were evaluated (Table 1). Both configurations were trained with the same set of active NES motifs that have been validated using the export assay and therefore, have been assigned an activity score. The activity score of each motif was taken into account to compute the PSSM in Wregex D (Supplementary Table S4), but not in Wregex C (Supplementary Table S3). As shown in Figure 6 and Table 4, Wregex D predicted a lower number of candidates than Wregex C (57 vs 83), but the number of true positives was also lower (22 vs 30). Thus, computation of the PSSM using the activity score leads to a more restrictive prediction, but does not appear to increase the proportion of true positive candidates. Therefore, we favor the option of computing the PSSM without taking into account the export assay activity scores. This would allow including a larger number of experimentally validated NES motifs (i.e. those not having an activity score) for PSSM computation.

### 3.4 Recommended configuration for searching NES motifs: Wregex E

Based on the results of the comparisons described above, we recommend a Wregex configuration for NES motif searching (Wregex E) that uses the WRE regular expression and a PSSM (Supplementary Table S5) computed using the experimentally validated NES motifs reported in ValidNESs (Fu et al., 2013), in the DUB NES survey of Garcia-Santisteban et al. (2012) and in this report. This configuration is implemented in the Wregex application webpage as “Recommended”, and an example of the results obtained is shown in Figure 7 and Table 4.

---

**Table 3. Result summary for Wregex B and ELM predictions.**

<table>
<thead>
<tr>
<th>Software</th>
<th>Wregex B (Th=50)</th>
<th>ELM (CRM1_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>False Positives</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Not Tested</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Candidates</td>
<td>16</td>
<td>21</td>
</tr>
</tbody>
</table>

**Table 4. Result summary for Wregex C, D and E predictions using Th=50.**

<table>
<thead>
<tr>
<th>Software</th>
<th>Wregex C</th>
<th>Wregex D</th>
<th>Wregex E</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>30</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>False Positives</td>
<td>20</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Not Tested</td>
<td>33</td>
<td>20</td>
<td>42</td>
</tr>
<tr>
<td>Candidates</td>
<td>83</td>
<td>57</td>
<td>98</td>
</tr>
</tbody>
</table>

---

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on July 13, 2015
Fig. 6. Comparison Wregex C versus Wregex D. Graphs show the score assigned by Wregex C (a) or Wregex D (b) to each predicted NES motif versus the activity score experimentally assigned to each motif using a nuclear export assay. Green circles to the right and the left of the threshold indicate true positives and true negatives, respectively. Red circles to the right and the left of the threshold indicate false positives and false negatives, respectively. Black circles represent non-tested (NT) candidates.

Fig. 7. Results of NES prediction using Wregex E. Graphs show the score assigned by Wregex E to each predicted NES motif versus the activity score experimentally assigned to each motif using a nuclear export assay. Green circles to the right and the left of the threshold indicate true positives and true negatives, respectively. Red circles to the right and the left of the threshold indicate false positives and false negatives, respectively. Black circles represent non-tested (NT) candidates.

4 CONCLUSION
We have developed and tested a novel approach for motif searching consisting on the combination of both a regular expression and a Position-Specific Scoring Matrix (PSSM). This approach, termed “weighted regular expressions” (Wregex), can be regarded as an evolution of the Eukaryote Linear Motif (ELM) resource (Gould et al., 2010). Importantly, the introduction of a PSSM provides a score that may help prioritizing candidates for experimental testing.

Our initial motivation to develop Wregex was the prediction of NES motifs. In this regard, Wregex compares well with the most recently developed NES predictor, NESsential (Fu et al., 2011), being faster, and offering a better compromise between the number of potential candidates to test and the number of true positives.

Although we have focused here on NES motif searching, Wregex is a generic tool that can be used for searching other functional protein motifs. Thanks to the novel use of groups in the regular expression, the PSSM can be computed based on patterns rather than on strict positions. As a consequence, there is no need to make use of the artificial concept of alignment gaps, which has no utility when considering the physical conformation of the protein motif. Wregex also offers the possibility of considering input motif weight when building the PSSM. This is useful when the motif has a quantifiable activity rather than a yes/no effect. If this is the case, scores derived from an activity assay can be used for building the PSSM. Another important feature of Wregex is its fast execution, which make this tool useful for large scale database search.

ACKNOWLEDGEMENT
We are grateful to Drs. Paul Horton and Szu-Chin Fu (University of Tokyo), for their help with the definition of the NESsential training set, and to Dr. Beric Henderson (University of Sydney), for providing the pRev(1.4)–GFP vector. We are grateful to Dr. Sonia Bañuelos (UPV/EHU) for her helpful comments and the staff of the High Resolution Microscopy Facility (Sgiker-UPV/EHU).

Funding: This work was supported by the Basque Country Government Department of Industry (SAIOTEK S-PE12UN005 to JAR) and the University of the Basque Country (UI11/20).

REFERENCES


