molecular systems biology

# The Multiple Specificity Landscape of Modular Peptide Recognition Domains

David Gfeller, Frank Butty, Marta Wierzbicka, Erik Verschueren, Peter Vanhee, Haiming Huang, Andreas Ernst, Nisa Dar, Igor Stagljar, Luis Serrano, Sachdev Sidhu, Gary Bader, Philip Kim

*Corresponding author: Philip Kim, University of Toronto*

**Review timeline:**

|  |  |
|---|---|
| Submission date: | 23 September 2010 |
| Editorial Decision: | 03 December 2010 |
| Revision received: | 09 February 2011 |
| Accepted: | 11 March 2011 |

## Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 03 December 2010

Thank you again for submitting your work to Molecular Systems Biology. First of all, I have to apologize for the delay in getting back to you. We have now finally heard back from two of the three referees whom we asked to evaluate your manuscript. Unfortunately, the third reviewer failed to return a report. Given the recommendations provided by the present reviewers, I prefer to make a decision now with the available reports rather than delaying the process further. As you will see from the reports below, the referees find the topic of your study of potential interest. However, they raise several concerns on your work, which should be convincingly addressed in a revision of this manuscript. In particular, the reviewers feel that it would be important to provide a more detailed comparison of your approach with other available methodologies (ANN, HMM or methods based on prior structural information).

We would also kindly ask you to deposit the validated protein interactions in an appropriate database from the International Molecular Exchange Consortium (IMEx, http://www.imexconsortium.org).

If you feel you can satisfactorily deal with these points and those listed by the referees, you may wish to submit a revised version of your manuscript. Please attach a covering letter giving details of the way in which you have handled each of the points raised by the referees.

I deeply apologize again for the lengthy process and look forward to receiving your revised manuscript.

Best wishes,

Editor

Molecular Systems Biology

---------------------------------------------------------------------------

REFEREE REPORTS:

Reviewer #1 (Remarks to the Author):

The binding of peptides to most adaptor domains, such as SH2, PDZ and WW, is known to occur through linear epitopes. Consequently, sequence-based analyses have tended to assume that the nature of amino acids at each position is uncorrelated. Kim and his colleagues have shown that this is untrue. They demonstrate through rigorous correlation analysis that occurrences of amino acids are very highly correlated between certain positions. They show that this can lead to classification of peptides binding SH3 domains into two groups. They further show that this corresponds to the two orientations of peptides binding, long ago characterized by structural biologists. Encouraged by this they have sought to correlate classifications of PDZ domains in terms of structural differences and achieved this in a convincing manner. The original focus on WW domains seems to be lost at this stage.

Nevertheless, this paper should be published. It brings together amino acid sequence and structural studies to give a better understanding of specificity. As such it informs the definition of interactions between pathways that are of central interest to molecular systems biology.

I would like to see some comment added in a revised version as to whether the sequence classes could have been predicted from substitution tables that take into account the local structural environment of each sidechain position of the already published structures. These they analyse retrospectively, even though they have been available for some time.

It would also be helpful to know if the correlation analyses can distinguish between correlations due to different binding modes (different orientations, different conformations of the binding modules) and those arising from tertiary interactions, i.e. sidechain-sidechain interactions within peptides.

Reviewer #2 (Remarks to the Author):

In this work the authors aim to better understand how modular protein domains such as WW, PDZ, SH2, SH3 may interact with the cognate ligands referred to as linear motifs.

In particular the authors investigate couplings between amino acids within the peptide bound by PDZ and other domains. They discuss how previous/existing computational approaches based on position specific scoring matrices do not capture such inter-residue correlations and thus under-perform in predicting or identifying such motifs.

Major Issues

1. While the authors do have a point, this is not all news. It is well known that Artificial Neural Networks and other statistical approaches such as Hidden Markov Models are capable of capturing

such correlations. In addition to this the authors claim that such models do not provide 'simple visualisation and direct interpretation', this is a somewhat flawed argument:

Firstly, such approaches can indeed be used to generate sequence logo's similar to those found in the present paper as such can be derived as linear approximations to a neural network. Secondly, this has already been done in e.g. the Miller et al. work (Science Signaling 2008) in which the authors looked at inter-residue couplings in linear motifs through training of linear and non-linear ANNs. The authors should for sure mention this work as a preceding one to their own work. In fact this is a main reason ANNs have a long track-record in dealing with linear motifs (e.g work by Brunak's lab).

2. The approach chosen is not discussed or validated against other approaches at a first glance more suitable approaches (such as ANNs or HMMs). It is well known that PSSMs or weight matrices can function as 'linear approximations' of ANNs, thus the use of multiple PWMs seems like an attempt to put in non-linearity into a model which is intrinsically linear. The referee would like to have this discussed much more in details in the paper as well as a comparison of performance/observations between the approach the authors are using and a more powerful non-linear model such as an ANN or HMM. The authors are commended from performing AROC analysis of their

performance in predicting physical interactions with their PWMs and they also validate some experimentally which enhances the impact of the paper. However, it would be very useful with a direct comparison with another computational approach which could give additional or novel insight not picked up by the PWM approach.

Specific Issues

1) As discussed above it is not correct that artificial neural networks (Brusic) or HMMs can not be visualised or interpreted easily. It may be defended this is less direct than in the case of a PWM but then again what the authors are trying to capture is inter-residue couplings which inherently would require a non-linear model such as a multi-layered ANN. This should be discussed in the introduction and end of the paper.

2) There seem to be no benchmarking of the multiple PWM method itself, compared to another approach. In addition the AROC analysis for prediction of physical interactions seem to not have P-values analysis performed to show that the AROCs are indeed statistically significant, this should be carried out.

3) The authors could discuss how their result may change if one were to distinguish substrate specificity from peptide specificity.

4). The authors should show the Mutual Information profile between all positions of DLG1 PDZ1 binding peptides in figure 1.

5) Further, in figure 1 it is not clear why the last sequence is not an independent cluster by itself? The choice of the number of clusters should be explained in the main text. Also how sensitive is the approach to the number of clusters? How does the performance change as a function of this?

6) The use of Gene Ontology does not add any value to the manuscript (on the contrary) the referee suggest to remove this part of the manuscript. It would be much more useful with more experimental validation.

7). There are some unclear parts of the text which could do with rewriting:

7.A) Some of the notation (e.g. Peptide recognition modules (PRMs), binders or ploytopic which I assume is polytopic) are not the terms used by the majority of the community. Modular Protein Domains is the predominant term for these protein domains and the polytopic concept is undefined and unexplained. This should be re-written.

7.B) There are some errors of punctuation and some extra references are needed in some places (e.g. WW domains data set?)

5.C) There are some contractions that should probably be avoided.

8. It could be interesting to quantify the difference between multiple PWMs (as introduced in the Discussion section) by changes in entropy or by relative entropy. The authors should at least briefly discuss this, better would be to include in the paper.

1st Revision - authors' response 09 February 2011

Reviewer Comment

I would like to see some comment added in a revised version as to whether the sequence classes could have been predicted from substitution tables that take into account the local structural environment of each sidechain position of the already published structures. These they analyse retrospectively, even though they have been available for some time.

Author Response

We have added a number of new calculations to see whether some of the new binding modes could have been predicted from structure alone. In short, while structural models can yield some information on binding specificities, they cannot predict new binding modes in absence of substantial amounts of data, most notably on the backbone conformations in the new binding mode. We have added this analysis into the manuscript.

Manuscript Changes

Added one paragraph at the end of section ëMultiple specificity predicts new binding modes of PDZ domains’. Added Figure S6. Added one paragraph in Supp info (ëStructural Analysis with FoldX’)

Reviewer Comment

It would also be helpful to know if the correlation analyses can distinguish between correlations due to different binding modes (different orientations, different conformations of the binding modules) and those arising from tertiary interactions, i.e. sidechain-sidechain interactions within peptides.

Author Response

We thank you for raising this interesting point. We have added a discussion about this in the paper and propose a simple method to distinguish between the two kinds of correlations. Briefly speaking, we use both the distance separating correlated positions and the difference in specificity at these positions. If correlated positions are all close to each other, we suggest that they correspond to tertiary interactions, while if some of the correlated positions are far from each other and display very different specificity in the different clusters, correlations are more likely to originate from different binding modes.

Manuscript Changes

Added in Discussion section: A possible way to automatically distinguish between the two kinds of multiple specificity is to compute the residue preference similarity between correlated positions. Based on our results, we suggest that if clear differences are found at three or fewer correlated positions, and all of them are comprised within 4 residues, then the correlations most likely correspond to interactions between ligand residues. Conversely, if all correlated positions are different (e.g., DLG1#1), or if correlated positions are far away from each other (e.g., SH3 domains in Figure 4), multiple specificity is more likely to correspond to distinct binding modes.

Reviewer Comment

While the authors do have a point, this is not all news. It is well known that Artificial Neural Networks and other statistical approaches such as Hidden Markov Models are capable of capturing such correlations. In addition to this the authors claim that such models do not provide 'simple visualisation and direct interpretation', this is a somewhat flawed argument:

Firstly, such approaches can indeed be used to generate sequence logo's similar to those found in the present paper as such can be derived as linear approximations to a neural network. Secondly, this has already been done in e.g. the Miller et al. work (Science Signaling 2008) in which the authors looked at inter-residue couplings in linear motifs through training of linear and non-linear ANNs. The authors should for sure mention this work as a preceding one to their own work. In fact this is a main reason ANNs have a long track-record in dealing with linear motifs (e.g work by Brunak's lab).

Author Response

We thank the referee for this insightful comment. ANNs and HMMs have indeed been applied to similar problems before, we had a brief discussion of such approaches in the intro of the paper, but have expanded it now considerably. We agree with the reviewer that ANNs can be used to derive a sequence logo, but for studies using ANNs, such visualization has not lead to an extensive mapping of multiple specificities and prediction of distinct binding modes, which is the focus of our work. We have included references of Miller 2008, Nielsen 1999 and Blom 1999 among others, as previous work that use other machine learning techniques.

Manuscript Changes

Introduction: Other machine learning algorithms, such as Hidden Markov Models (HMM) (Noguchi et al, 2002) or Artificial Neural Networks (ANN) (Blom et al, 1999) (Brusic et al, 1998) (Emanuelsson et al, 2000) (Miller et al, 2008) have been used previously in different contexts to account for positional correlations (Nielsen et al, 1999). [...] Moreover, thanks to simple visualization (which, mathematically speaking, can be related to linear approximations of HMMs or ANNs) and direct interpretation, the multiple specificity model reveals new structural insights into binding modes of Modular Protein Domains and predict new protein interactions within signaling pathways mediated by these domains.

Reviewer Comment

The approach chosen is not discussed or validated against other approaches at a first glance more suitable approaches (such as ANNs or HMMs). It is well known that PSSMs or weight matrices can function as 'linear approximations' of ANNs, thus the use of multiple PWMs seems like an attempt to put in non-linearity into a model which is intrinsically linear. The referee would like to have this discussed much more in details in the paper as well as a comparison of performance/observations between the approach the authors are using and a more powerful non-linear model such as an ANN or HMM. The authors are commended from performing AROC analysis of their performance in predicting physical interactions with their PWMs and they also validate some experimentallywhich enhances the impact of the paper. However, it would be very useful with a direct comparison with another computational approach which could give additional or novel insight not picked up by the PWM approach.

Author Response

We have added a detailed comparison with ANNs and HMMs using ROC curves. The performance is indeed very similar. We agree that multiple PWMs may be obtained as ëlinear approximations' of more complex models (although, mathematically speaking, the mixture of PWMs model is not linear), and we've included this point in the introduction. The main benefit of our work remains to show that a relatively simple and intuitive framework canaccurately model positional correlations and enabled us to reveal multiple specificities and predict new binding modes.

Manuscript Changes

Table S1 and and S4 for the AROC values. Section "Multiple PWMs more accurately model binding specificity" has been changed to include comparison with HMMs and ANNs both in terms of cross-validation and testing with independent datasets. The full description of HMMs and ANNs used in this work has been included in Supplementary information.

Reviewer Comment

1) As discussed above it is not correct that artificial neural networks (Brusic) or HMMs can not be visualised or interpreted easily. It may be defended this is less direct than in the case of a PWM but then again what the authors are trying to capture is inter-residue couplings which inherently would require a non-linear model such as a multi-layered ANN. This should be discussed in the introduction and end of the paper.

Author Response

We apologize for not mentioning the linear approximation of HMM and ANN. We've included this point in the introduction and discussion. Along this line, we've observed that most of the HMMs models did converge to a transition state diagram with n almost independent paths, where n is the number of different PWMs, further highlighting that in our case, the multiple PWMs accurately capture positional correlations.

Manuscript Changes

Several part of the introduction (see above). Discussion: From a computational point of view, the multiple PWMs give similar performance as other machine learning algorithms. We suggest that, because of the structural constraints underlying short peptide binding events, a simple decomposition into multiple PWMs is sufficient to handle correlations, while more complex processes, such as sub-cellular localization (Emanuelsson et al, 2000), require more advanced machine learning algorithms.

Reviewer Comment

2) There seem to be no benchmarking of the multiple PWM method itself, compared to another approach. In addition the AROC analysis for prediction of physical interactions seem to not have P-values analysis performed to show that

the AROCs are indeed statistically significant, this should be carried out.

Author Response

We've added a full benchmarking of the multiple PWMs model both using cross-validation and testing with independent datasets, and comparison with HMMs and ANNs. The AROC p-values have been included in Table S4, showing that most AROCs are indeed statistically significant.

Manuscript Changes

Table S1 and and S4. Section "Multiple PWMs more accurately model binding specificity" has been significantly changed

Reviewer Comment

3) The authors could discuss how their result may change if one were to distinguish substrate specificity from peptide specificity.

Author Response

In this work we've focused on Modular Protein Domains interacting with short linear peptides. For interactions involving larger substrate (for instance other domains), the correlation patterns may be much more complex and it may be difficult to define different binding modes. We've included this point in the discussion.

Manuscript Changes

Added in Discussion: For other kind of protein interactions, such as the ones involving larger binding interface or non-peptide substrates, sequence-based approaches are more difficult to apply. As such, the multiple PWMs model is especially suited for proteins interacting with small ligands made out of a limited number of building blocks (e.g., amino acids or nucleotides) and adopting a few different binding modes on their targets.

Reviewer Comment

4). The authors should show the Mutual Information profile between all positions of DLG1 PDZ1 binding peptides in figure 1.

Author Response

This has been included in Figure 1 on top of panel A

Manuscript Changes

Figure 1A

Reviewer Comment

5) Further, in figure 1 it is not clear why the last sequence is not an independent cluster by itself? The choice of the number of clusters should be explained in the main text. Also how sensitive is the approach to the number of clusters? How does the performance change as a function of this?

Author Response

We've included the description of the choice of the number of clusters in the main text. This number is chosen as the minimal number of clusters such that no positional correlations are observed within each cluster. For instance in Figure 1, the peptides in each of the two clusters do not display internal correlations according to the p-value threshold, so there is no need of further splitting. The last sequence, although shorter than the other ones in the cluster, seems to follow a similar binding mode with W at position 0, a hydrophobic residue at -1 and a polar residue (S/T) at -3.

We've tried to vary the number of clusters, for instance by adding one cluster (i.e. one component in the mixture model). Overall, the performance in terms of cross-validation AROC values are similar (see Table S1). However, in more than one third of the cases, the mixture model was automatically giving a weight lower than 0.05 to one of the components, suggesting that this additional cluster was not very relevant. Moreover, in many other cases, the multiple logos become quite redundant and do not seem to lead to additional insights.

Manuscript Changes

Added: The main idea of this approach is to fit K different PWMs to the aligned peptides, where K is chosen here as the number of clusters found to remove positional correlations. Table S1: results of the multiple PWMs with one more cluster.

Added in Supplementary information: We also tried to change the value of K, to explore the sensitivity of the model with respect to this parameter. Adding one cluster for each domain, we observed that, in 38% of the cases, one of the components of the mixture model was given a weight lower than 0.05, suggesting that this additional component is not relevant. In many other cases some of the multiple PWMs were clearly redundant. In terms of cross-validation performance (see Table S1), we did not observe significant changes.

Reviewer Comment

6) The use of Gene Ontology does not add any value to the manuscript (on the contrary) the referee suggest to remove this part of the manuscript. It would be much more useful with more experimental validation.

Author Response

This part has been removed from the manuscript.

Reviewer Comment

7.A) Some of the notation (e.g. Peptide recognition modules (PRMs), binders or ploytopic which I assume is polytopic) are not the terms used by the majority of the community. Modular Protein Domains is the predominant term for these protein domains and the polytopic concept is undefined and unexplained. This should be re-written.

Author Response

Peptide Recognition Module has been replaced and the abbreviation has been removed. We have used Modular Protein Domains to refer to general protein domains and Modular Peptide Recognition Domains when it was necessary to specify that our results apply to modular protein domains interacting with short peptides. To simplify the manuscript, we removed the polytopic term, and simply explain that we focus on Modular Protein Domains with a single binding site and known to interact with short linear peptides.

Reviewer Comment

7.B) There are some errors of&#x00A0;punctuation&#x00A0;and some extra references are needed in some places (e.g. WW domains data set?)

Author Response

The full dataset for the three WW domains, which has been generated by ourselves for the purpose of this study, has been added as Supplementary Table S5. Many references have been included, especially regarding different machine learning algorithms.

Reviewer Comment

5.C) There are some contractions that should probably be avoided.

Author Response

We’ve tried to improve this by keeping only the contractions that are extensively used throughout the manuscript.

Reviewer Comment

8. It could be interesting to quantify the difference between multiple PWMs (as introduced in the Discussion section) by changes in entropy or by relative entropy. The authors should at least briefly discuss this, better would be to include in the paper.

Author Response

We thank you for raising this interesting point. We've included a paragraph and a Figure (Figure S2) showing how considering multiple PWMs leads to a statistically significant decrease in entropy (P<0.001). In particular, several domains seem to have an almost flat profile when modeled with a single PWM, while our multiple PWMs analysis reveals that there are in fact highly specific in their multiple specificity profiles.

Manuscript Changes

Figure S2 and added at the end of section ëPositional correlations originate from multiple specificity': It is also interesting to observe that clustering the peptides lead to an enhanced specificity, with an average entropy over all positions and all domains of 0.52 before clustering and 0.42 after clustering (P<10-4 when compared with random clusters, see Supplementary information and Supplementary Figure S2). In particular multiple PWMs reveal interesting specificities that tend to be smoothed out in the single PWM visualization (see for instance MLLT4#1 or HTRA2#1 in Figure 2).