



- Article -

Improved prediction of allergenicity by combination of multiple sequence motifs

Waiming Kong*, Tsu Soo Tan, Lawrence Tham and Keng Wah Choo

Bioinformatics Group, Nanyang Polytechnic, 180 Ang Mo Kio Ave 8, Singapore (589 830)

* Corresponding author

Email: KONG_Wai_Ming@nyp.gov.sg

Edited by H. Michael; received June 27, 2006; revised December 06, 2006; accepted December 07, 2006; published January 05, 2007

Abstract

The identification and validation of protein allergens have become more important nowadays as more and more transgenic proteins are introduced into our food chains. Current allergen prediction algorithms focus on the identification of single motif or single allergen peptide for allergen detection. However, an analysis of the 575 allergen dataset shows that most allergens contain multiple motifs. Here, we present a novel algorithm that detects allergen by making use of combinations of motifs. Sensitivity of 0.772 and specificity of 0.904 were achieved by the proposed algorithm to predict allergen. The specificity of the proposed approach is found to be significantly higher than traditional single motif approaches. The high specificity of the proposed algorithm is useful in filtering out false positives, especially when laboratory resources are limited.

Keywords: allergen, allergen prediction, combination of motifs

Introduction

Allergens induce allergic responses by the elicitation of immunoglobulin E (IgE) antibodies, causing the numerous symptoms of allergies within affected individuals. Allergies affect many people and are among the most common causes of chronic illnesses in developed countries. Although the advancement in pharmaceutical therapies including antihistamines, corticosteroids and injectable epinephrine help alleviate the symptoms of

TABLE of CONTENTS:

Title

Abstract

Introduction

Materials and methods

Results

Comparing motif combination method with single motif detection method

Comparing combination of motifs method with other prediction methods

Comparing combination of motifs method with other prediction methods using the Swiss-Prot database

Comparison of predicted results with experimental results

Conclusion

References

Figures

1 2 3 4 5 6

Tables

the allergies, avoidance of the offending allergen is deemed the most effective treatment, if the offending allergen is known to the affected individual. Generally, allergen avoidance is more effective when the offending allergens are proteins found in our dietary food as the allergenic protein can be avoided by abstinence of food sources containing the offending protein. The identification and validation of food protein allergens is therefore of utmost importance in the study of allergies, more so as more and more transgenic proteins are introduced into our food chains.

Laboratory assessment of protein allergenic potential, including immunogenicity, cross-reactivity and clinical symptoms studies are both time-consuming and cumbersome. While animal studies of immune responses, by means of measuring immunoglobulin E (IgE) antibody or T-cell response, provide the ultimate validation of an allergen, they are too expensive to be applied to every protein. Therefore, a reliable prediction of potential allergens is imperative for allocating the money to study the real potential allergens, rather than all proteins.

It is well known that IgE and T-cell responses are epitope-dependent [Jahn-Schmid *et al.*, 2002]. The study of hundreds of allergen protein sequences suggest that allergens tend to share certain sequence similarities. Thus, the potential allergenicity of query proteins can be predicted by examining their sequence similarities with known allergens. Attempts [Fiers *et al.*, 2004; Hileman *et al.*, 2002; Zorzet *et al.*, 2002] have been made to predict allergenicity of a query protein by its amino acid sequences. According to the FAO/WHO guidelines [FAO/WHO, 2001] for allergenicity evaluation of foods derived from biotechnology, a query protein is potentially allergenic if it either has an identity of at least six contiguous amino acids (identity length $n = 6$) with an known allergen or >35% sequence similarity over a window of 80 amino acids when compared with known allergens. In 2003, the Codex Alimentarius [Codex Alimentarius Commission, 2003] recommended that the size of the continuous amino acid search should be based on a scientifically justified rationale in order to minimize the potential for false negative or false positive results. In an attempt to improve the prediction efficacy, Stadler and Stadler, 2003, generated a minimal set of motif sequences from a database of known allergens, classifying unknown proteins as potential allergens if the proteins contain at least one of the motif sequence generated. The specificity of this method showed a noticeable improvement over the FAO/WHO guidelines. Li *et al.*, 2004, further refined the aforementioned method by first clustering the sequences of known allergens before generating conserved motifs for each cluster. A profile for each predicted motif is built and used for predicting the allergenicity of a query protein sequence. Björklund *et al.*, 2005, on the other hand, used allergen-representative peptides (ARP) to predict allergen sequences. Allergen sequences were first cut into short fixed-length peptides. Similarity scores with a database of 21476 non-allergen peptides were calculated for each fixed-length peptide. Peptides with low similarity scores were selected to form the ARP dataset that was used for allergen prediction. This method achieved very high prediction rates. Further improvements to the ARP method were made by Soeria-Atmadja *et al.*, 2006, through the use of the variable-length peptides. The authors used Support Vector Machine (SVM) to classify variable-length peptides to achieve greater specificity than the ARP method. A hybrid prediction method that combined SVM and IgE epitope-based predictions, proposed by Saha *et al.*, 2006, achieved a commendable accuracy of 85%. The authors set up the AlgPred web server that allows prediction of allergens from amino acid sequence of the protein using any of the following approaches; (i) scanning of IgE epitopes; (ii) motif-based approach; (iii) SVM-based method using amino acid composition of protein; (iv) Hybrid approach; and (v) BLAST search on ARPs.

Current allergen prediction algorithms [Kleter *et al.*, 2002; Stadler and Stadler, 2003; Li *et al.*, 2004] are often based on the association of unknown sequence with a single allergen motif or an allergen peptide. However, it was observed from biological experiments [Kurup *et al.*, 1998;

[Kurup et al., 2003](#); [Ramos et al., 2003](#)] that allergens usually contain multiple epitopes. In the case of discontinuous epitopes [[Ganglberger et al., 2000](#); [Ramachandran et al., 2002](#)], epitopes are formed by fragments that are brought together by the physical bending of the protein chain. Through the studies of PDB structures [Dall'Acqua et al., 1996](#), and [Mirza et al., 2000](#), showed that allergens were bound to antibody through 2 interaction sites. The authors of this paper first investigated on the number of motifs found in allergen sequences. Based on the findings, a new allergen prediction algorithm was devised based on the combination of motifs, rather than single motif. The proposed algorithm predicted allergenicity of proteins with a sensitivity of 0.772 and specificity of 0.904, markedly reducing false positive rates as compared to algorithms that focuses on one motif or epitope.

Materials and methods

Allergen and non-allergen data

Allergen data was obtained from the allergen list published by [Bjorklund et al., 2005](#). We chose to use this database (575 proteins) as each record was manually inspected for documentation on allergenicity. In this database, sequences with less than 100 amino acids in length and records with poor documentation were removed. The non-allergen test set of 700 sequences was obtained from the same source. The Swiss-Prot database (release 50.9, 17 October 06), consisting of 235673 protein sequences, was used for the evaluation comparison with other prediction methods in section "[Comparing combination of motifs method with other prediction methods using the Swissprot database](#)".

Allergen prediction by combination of motifs

Currently available allergen prediction algorithms [[Stadler and Stadler, 2003](#); [Li et al., 2004](#)] are based on sequence similarity with a single motif. When motifs are extracted from the allergen dataset using the MEME (Multiple Expectation Maximization for Motif Elicitation) software [[Bailey and Elkan, 1994](#)], it was discovered that most of the allergens contain multiple motifs as shown in [Fig. 1](#). It was observed that 98.6% of all the allergens contain a minimum of 2 motifs and 35.8% contain 4 or 5 motifs. Based on the aforementioned finding, instead of using a single motif for allergen detection, our allergen prediction method utilizes a combination of motifs to improve allergen prediction. As will be demonstrated below, the use of a combination of 2 motifs for allergen detection yielded higher sensitivity and specificity than currently available allergen predictions based on single motif detection.

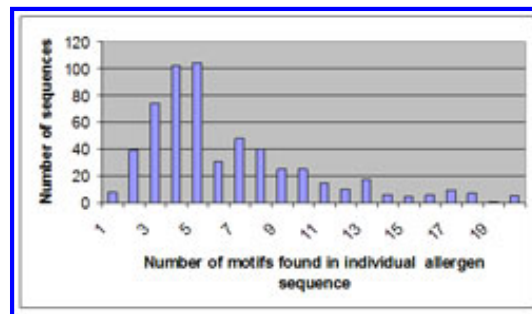


Figure 1: Histogram showing the distribution of the number of motifs extracted from the allergen sequences. 98.6% of the allergen dataset (575 sequences) contain 2 or more motifs.

Allergen prediction method

The proposed allergen prediction method relies primarily on the detection of a combination of 2 allergen motifs in a given protein sequence. From the positive training data set mentioned in section "[Allergen and non-allergen data](#)", a set of allergenic motifs (N) was generated by the MEME software [[Bailey and Elkan, 1994](#)]. A database (D) containing all possible combinations of 2 motifs from the set of allergenic motifs (N) was then generated. In the example of an allergenic protein containing 3 motifs ([Fig. 2](#)), M1, M2 and M3, the possible combination sets [M1, M2], [M1, M3] and [M2,M3] were generated and stored in database (D).

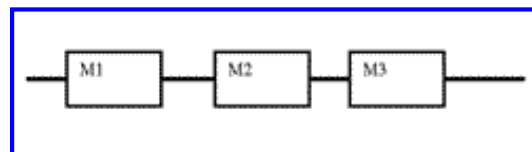


Figure 2: Motifs M1, M2, M3 found in an allergenic protein.

Using the motifs generated by MEME (N), Motif Alignment and Search Tool (MAST) [[Bailey and Gribskov, 1998](#)] was used to ferret out motifs in any given protein sequence. MAST is a tool for identifying a set of given motifs from any given protein sequence. MAST takes in the output file of MEME, containing the descriptions of one or more motifs, as the input file and searches any query protein for matches to the motifs that is found in the input file. The motifs identified by MAST, using the default parameters, in the query protein sequence were then compared with the motif combination database (D). In the example of a query protein containing 2 motifs (R1and R2), the combination set [R1, R2] will be used to search the motif database. Detection of the exact combination of motifs in the database renders the protein sequence the status of 'potentially allergenic' by the proposed allergen detection algorithm.

Accuracy of diagnostic tests

Accuracy of a diagnostic test can be expressed through sensitivity and specificity. The calculations for sensitivity and specificity are as follows:

Sensitivity = $TP/(TP + FN)$,

Specificity = $TN/(TN + FP)$,

where TP is the number of true positives, FP = number of false positives, FN = number of false negatives, TN = number of true negatives.

Five-fold cross-validation

Five-fold cross-validation was used to verify the effectiveness of the allergen prediction. The procedure is as follows:

- 1. The positive allergen data was divided into 5 sets, each containing 115 allergen sequences.
- 2. One set of data was used as the positive test data while the remaining 4 sets were combined to form the training data set. The negative data set consists of the 700 non-allergen sequences as described in section " Allergen and non-allergen data".
- 3. Specificity and sensitivity tests were measure and recorded using the proposed allergen prediction method described in section "Allergen prediction method".
- 4. Steps 2 and 3 were repeated five times, using a different set of data from the 5 sets of positive allergen data as the positive test data.

Results

The five-fold cross-validation described in section "Five-fold cross-validation" employing the proposed allergen prediction method described in section "Allergen prediction method" was performed on different motif lengths and motif numbers using the MEME program. Four motif lengths (5 amino acids [aa], 15 aa, 25 aa, 35 aa), and five different motif numbers (100, 200, 300, 400, 500) were tested.

The sensitivity results for the different motif lengths and motif numbers are given in Tab. 1. Fig. 3 shows the plots for the medium values of sensitivity, specificity versus motif length. From the plot, it was observed that as the motif length increases, sensitivity increases. This observation concurred with Hudeez's findings that the minimum peptide length needed to bind initiate a T-cell response is between 7 and 8 residues [Hudeez, 1994]. Specificity also increases as the motif length increases.

Table 1: Sensitivity and specificity results.

Motif length (aa)	Number of motifs	Sensitivity	Specificity
	100	0.372	1.000
	200	0.336	0.809

5	300	0.311	0.777
	400	0.318	0.752
	500	0.311	0.750
15	100	0.600	0.950
	200	0.697	0.885
	300	0.704	0.827
	400	0.723	0.805
	500	0.697	0.789
25	100	0.649	0.926
	200	0.736	0.869
	300	0.758	0.855
	400	0.748	0.841
	500	0.760	0.846
35	100	0.633	1.000
	200	0.708	0.870
	300	0.753	0.873
	400	0.755	0.909
	500	0.772	0.904

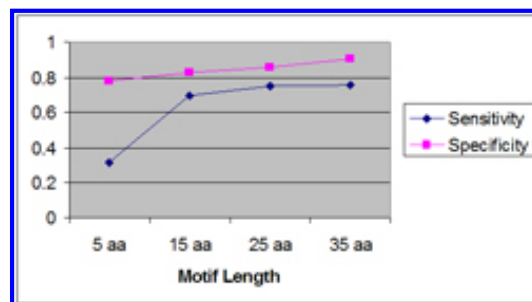


Figure 3: Plots of sensitivity, specificity vs motif length.

Fig. 4 is a plot of the medium values of the sensitivity and specificity versus the motif numbers used for prediction. From Fig. 4, it can be seen that as the number of motifs increases, sensitivity increases. However, for number of motifs greater than 300 motifs, increase in numbers of motifs does not lead to any significant increase in sensitivity. On the other hand, specificity decreases as the number of motifs increases. This may be due to the fact that with fewer motifs, the algorithm picks up less false positive, resulting in high specificity. From Fig. 4, we can conclude that the most important motifs contributing to allergen prediction are contained in the initial 300 motifs ferreted by MEME.

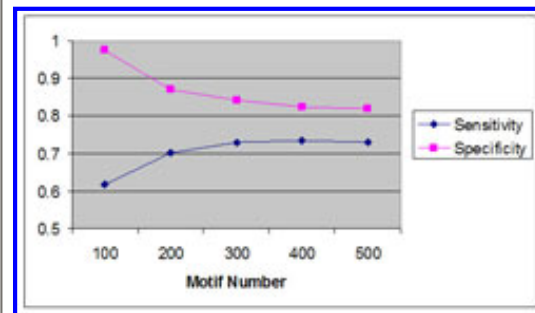


Figure 4: Plots of sensitivity, specificity vs motif number.

In summary, both sensitivity and specificity increase with longer motif length. However, in the case of motif number, sensitivity increases while specificity decreases when the motif number increases. In conclusion, a good allergen predictor should have long motif length to achieve high sensitivity and specificity. The motif number, on the other hand, should be chosen carefully to strike a balance between sensitivity and specificity. From our experiments, as shown in Figs. 3 and 4, we recommend motif length of at least 25 aa and motif numbers between 300 and 500 for optimal sensitivity and specificity.

Comparing motif combination method with single motif detection method

Current available allergen detection algorithms focus on the use of single motif or allergen peptide. In this paper, the authors compare the prediction results of using a combination of motifs against prediction results of using a single motif. We implemented the single motif prediction similar to the steps of employing a combination of motifs with the exception of generating a motif combination database. In the case of a single motif method, a database of all motifs was used instead of the motif combination database. For the single motif test, protein sequences with any of the motifs found in the single motif database were predicted as allergenic.

The specificity and sensitivity of the prediction results for the combination of motifs method versus the single motif method are shown in Figs. 5 and 6. From Fig. 5, it can be seen that the sensitivity of combination of motifs method and single motif method are very similar, differing only by

about 0.7-3.65% for the different motif lengths (Tab. 2). However, the specificity of combination of motifs method performed much better than the single motif method, out performing it by about 5.5 to 13.0%.

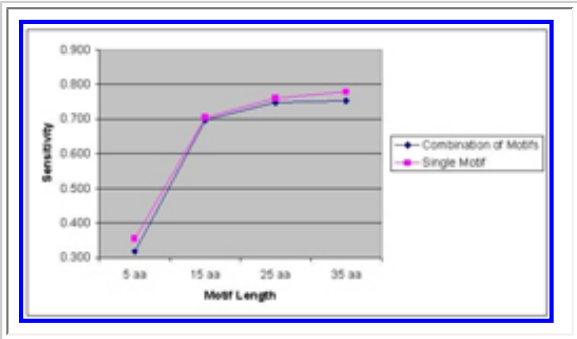


Figure 5: Plot of sensitivity vs motif length for combination of motifs and single motif algorithms.

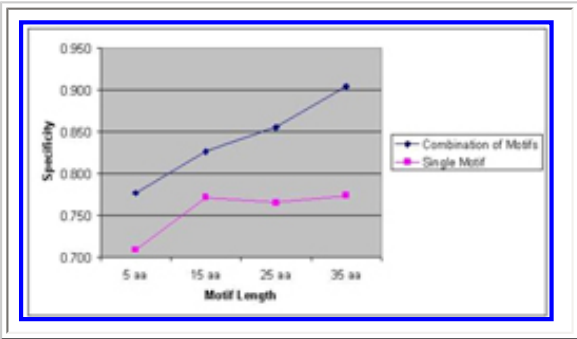


Figure 6: Plot of specificity vs motif length for combination of motifs and single motif algorithms.

Table 2: Medium sensitivity and specificity results for different motif lengths.

Prediction method	Motif length (aa)	Sensitivity	Specificity
Combination of motifs	5	0.318	0.777
	15	0.697	0.827
	25	0.748	0.855
	35	0.753	0.904
Single motif	5	0.355	0.709
	15	0.704	0.772

	25	0.760	0.765
	35	0.779	0.774

Comparing combination of motifs method with other prediction methods

Results of the proposed combination of motifs method were also compared with other prediction methods. The prediction results of other methods are obtained from the work published by [Saha et al., 2006](#). We believe that the comparison is valid as the same datasets and five-fold cross-validation tests were conducted for all the prediction methods. From [Tab. 3](#), it was observed that the IgE method performed poorly, achieving sensitivity result of only 0.157. However, the sensitivity of the IgE method could improve if more IgE epitope data was available. The Mast prediction method, on the other hand, could only achieve either high sensitivity or specificity, but not both, as shown in [Tab. 3](#) for Mast prediction method with *E*-value of 0.1 and Mast prediction method with *E*-value of 100. Only the results of the SVM and hybrid methods were comparable to the proposed combination of motifs method. Although the SVM and the hybrid methods yielded better sensitivity, the combination of motifs method produced better specificity. It is observed from [Tab. 3](#) that the Blast(ARP) method provides the best results among all the predictors. The use of a large non-allergen training set helps to give this method superior performance. From [Tab. 3](#), it can be seen that the combination of motifs method ranks among the top allergen predictors that can achieve good sensitivity and specificity.

Table 3: Comparing combination of motifs prediction results with other prediction methods.

Prediction methods	Sensitivity	Specificity
Mast(ev0.1)	0.149	0.911
Mast(ev100)	0.939	0.331
IgE(PID876)	0.157	0.984
Combination of motifs (25 aa 300 motifs)	0.772	0.904
SVM	0.889	0.819
Hybrid	0.889	0.819
Blast(ARP)	0.836	0.979

Comparing combination of motifs method with other prediction methods using the Swiss-Prot database

Comparison with other prediction algorithms using the entire Swiss-Prot database (release 50.9, 17 October 06, 235673 sequences) was performed. The DFLAP prediction result was obtained from work by [Soeria-Atmadja et al., 2006](#), and the rest of the results were obtained from the paper by [Saha et al., 2006](#). The combination of motifs method achieved a high specificity of 0.9628, better than the SVM and the Mast(ev100) methods. Although the IgE and the Mast (ev0.1) prediction methods could achieve better specificity results ([Tab. 4](#)), the sensitivity of these methods were poor, as shown in [Tab. 3](#). The Blast(ARP) and the DFLAP methods which used a large non-allergen database for training are currently the best allergen predictors available.

Table 4: Comparing combination of motifs specificity results with other methods using the Swiss-Prot database.

Prediction methods	Specificity
SVMc	56.07
SVMd	61.09
Mast (ev100)	86.68
Combination of motifs	96.28
Mast(ev0.1)	96.58
Blast(ARP)	97.97
IgE epitope	98.25
DFLAP	98.5

Comparison of predicted results with experimental results

The proposed algorithm was validated using allergens with known epitopes. Experimental data was obtained from the [SDAP database](#) (SDAP). In the case of the Par j 1 allergen, experimental data showed the presence of 4 discontinuous epitopes. The proposed algorithm predicted epitopes that covered the regions of the 4 true motifs as shown in [Table 5](#).

Table 5: Comparing experimental data of Par j 1 allergen with predicted epitopes using parameters motif length = 25 aa, motif number = 300. Predicted epitopes are highlighted in red and underlined while actual epitopes are highlighted blue and underlined.

Predicted Result (pos 1:70)	Q <u><u>ETCGTMVRALMPCLPFVQGKEKEPS</u></u> KGCCSGAKRLDGETKTGPQRVHACECIQTAMKTYSDIDGKLVSE
Experimental Data (pos 1:70)	QETCGTMVRALMPCLPF <u><u>VQGKEKEPS</u></u> KGCCSGAKRLDGET <u><u>KTGPQRV</u></u> HACECIQTAMKTYSDIDGKLVSE
Predicted Result (71:139)	V <u><u>PKHCGIVDSKLPPIDVNMDCKTVGV</u></u> VPRQPQLPVSLRH <u><u>GPVTGPS</u></u> PAHKARLER <u><u>PQIRVPPP</u></u> APEKA
Experimental Data (71:139)	V <u><u>PKHCGIVDS</u></u> SKLPPIDVNMDCKTVGVVPRQPQLPVSLRHGPVTGPS <u><u>DPAHKARLE</u></u> RPQIRVPPPAPEKA

For the allergen Jun a 3, experimental data showed the presence of 4 epitopes. Using parameters of motif length = 25 aa, motif number = 300, the proposed algorithm predicted 5 motifs which covered most of the regions of the 4 true epitopes. However, the first and the fifth predicted motifs were false positives.

Table 6: Comparing experimental data of Jun a 3 allergen with predicted epitopes using parameters motif length = 25 aa, motif number = 300. Predicted epitopes are highlighted in red and underlined while actual epitopes are highlighted blue and underlined.

Predicted Result(pos 1:70)	MAR <u><u>VSELAFLLAATLAISLHMQEAG</u></u> VVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNL <u><u>AAGTASARE</u></u>
Experimental Data (pos 1:70)	MARVSELAFLLAATLAISLHMQEAGVVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNL <u><u>AAGTASARE</u></u>
Predicted Result (71:140)	<u><u>WGRTGCTFDASGKGSC</u></u> QTGDCGGQLSCTV <u><u>SGAVPATLAEY</u></u> TQSDQDYDVSLVDGFNIPLAINPTNAQCT
Experimental Data (71:140)	<u><u>WGRTGCTFDASGKGSC</u></u> QTGDCGGQLSCTVSGAVPATLAEYTDQSDQDYDVSLVDGFNIPLAINPTNAQCT
Predicted Result (141:210)	APACKADINAVCPSELKV <u><u>DGGCNSACNVFKTDQYCCRNAYVDN</u></u> CPATNYSKI <u><u>FKNQCPQAYS</u></u> YAKDDTAT
Experimental Data (141:210)	APACK <u><u>ADINAVCPSELKVDGGCNSACNVFKT</u></u> DQYCCRNAYVDNCPATNYSKIFKNQCPQAYSYAKDDTAT
Predicted Result (211:225)	<u><u>FACASGT</u></u> DYSIVFCP
Experimental Data (211:225)	FACASGTDYSIVFCP

Conclusion

For allergen prediction, sensitivity is deemed more important than specificity as the possible consequences for missing a true allergen is more severe than producing a few false positive predictions. Nevertheless, a good allergen prediction method should possess both high sensitivity and

specificity. To achieve high sensitivity and specificity, we have proposed an algorithm to predict allergenicity of proteins using the combination of motifs found in known allergens. Using motif length of 35 amino acids and motif number of 500, the algorithm can achieve a sensitivity of 0.772 and a specificity of 0.904. Compared with previous approaches using single motif [Stadler and Stadler, 2003; Li *et al.*, 2004], we found that by employing a combination of motifs instead of a single motif, allergen prediction specificity increases significantly, while losing marginally in sensitivity. As such, the proposed combination motif method can be used to screen out maximum number of false positive in the situation where laboratory resources are limited. The method can be modified to use combinations of motifs that are extracted from different allergen sequences. This will enable the predictor to detect "new" allergens that contain motifs from different known allergens.

References

- Bufer, A. (2001). Significance of IgE-binding epitopes in allergic disease. *J. Allergy Clin. Immunol.* **107**, 219-221.
- Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28-36.
- Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54.
- Björklund, Å. K., Soeria-Atmadja, D., Zorzet, A., Hammerling, U. and Gustafsson, M. G. (2005). Supervised identification of allergen-representative peptides for *in silico* detection of potentially allergenic proteins. *Bioinformatics* **21**, 39-50.
- Dall'Acqua, W., Goldman, E. R., Eisenstein, E. and Mariuzza, R. A. (1996). A mutational analysis of binding of two different proteins to the same antibody. *Biochemistry* **35**, 9667-9676.
- FAO/WHO (2001). Allergenicity of genetically modified foods.
- Codex Alimentarius Commission. (2004) Foods Derived from Biotechnology. Joint FAO/WHO Food Standards Programme, Food and Agriculture Organisation of the UN WHO, Rome, Italy.
- Fiers, M. W. E. J., Kleter, G. A., Nijland, H., Peijnenburg, A. A. C. M., Nap, J. P. and van Ham, R. C. H. J. (2004). AllermatchTM, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* **5**, 133.
- Ganglberger, E., Grünberger, K., Sponer, B., Radauer, C., Breiteneder, H., Boltz-Nitulescu, G., Scheiner, O. and Jecsen-Jarolim, E.

- (2000). Allergen mimotopes for 3-dimensional epitope search and induction of antibodies inhibiting human IgE. *FASEB J.* **14**, 2177 - 2184.
-
- Hileman, R. E., Silvanovich, A., Goodman, R. E., Rice, E. A., Hollescha, G., Astwood, J. D. and Hefle, S. L. (2002). Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* **128**, 280-291.
-
- Hudeez, F. (1994). Prediction of B- and T-cell epitopes. *In: Synthetic peptides in the search for B- and T-cell epitopes*, Raynavolgyi, E. (ed.), R. G. Landis Company, Austin, TX, pp. 19-30.
-
- Jahn-Schmid, B., Kelemen, P., Himly, P., Bohle, B., Fischer, G., Ferreira, F. and Ebner, C. (2002). The T cell response to Art v 1, the major mugwort pollen allergen, is dominated by one epitope, *J. Immunol.* **169**, 6005-6011.
-
- Kleter, G. A. and Peijnenburg, A. A. C. M. (2002). Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE - binding linear epitopes of allergens. *BMC Struct. Biol.* **2**, 8.
-
- Li, K.-B., Issac, P. and Krishnan, A. (2004). Predicting allergenic proteins using wavelet transform. *Bioinformatics* **20**, 2572-2578.
-
- Kurup, V. P., Banerjee, B., Murali, P. S., Greenberger, P. A., Krishnan, M., Hari, V. and Fink, J. N. (1998). Immunodominant peptide epitopes of allergen, Asp f 1 from the fungus *Aspergillus fumigatus*. *Peptides* **19**, 1469-1477.
-
- Kurup, V. P., Vijay, H. M., Kumar, V., Castillo, L. and Elms, N. (2003). IgE binding synthetic peptides of Alt a 1, a major allergen of *Alternaria alternata*. *Peptides* **24**, 179-185.
-
- Mirza, O., Henriksen, A., Ipsen, H., Larsen, J., Wissenbach, M., Spangfort, M. and Gajhede, M. (2000) Dominant epitopes and allergic cross-reactivity: Complex formation between a fab fragment of a monoclonal murine IgG antibody and the major allergen from Birch Pollen Bet V 1. *J. Immunol.* **165**, 331-338.
-
- Ramachandran, H., Jayaraman, V., Banerjee, B., Greenberger, P. A., Kelly, K. J., Fink, J. N. and Kurup, V. P. (2002). IgE binding conformational epitopes of Asp f 3, a major allergen of *Aspergillus fumigatus*. *Clin. Immunol.* **103**, 324-333.
-
- Ramos, J. D., Cheong, N., Lee, B. W. and Chua, K. Y. (2003). Peptide mapping of immunoglobulin E and immunoglobulin G immunodominant epitopes of an allergenic *Blomia tropicalis* paramyosin, Blo t 11. *Clin. Exp. Allergy* **33**, 511-517.
-
- Saha, S. and Raghava, G. P. S. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, W202-W209.
-
- Stadler, M. B. and Stadler, B. M. (2003). Allergenicity prediction by protein sequence. *FASEB J.* **17**, 1141-1143.
-

- Saha, S. and Raghava, G. P. S. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, W202-W209.
- Stadler, M. B. and Stadler, B. M. (2003). Allergenicity prediction by protein sequence. *FASEB J.* **17**, 1141-1143.
- Soeria-Atmadja, D., Lundell, T., Gustafsson, M. G. and Hammerling, U. (2006). Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning. *Nucleic Acids Res.* **34**, 3779-3793.
- Zorzet, A., Gustafsson, M. and Hammerling, U. (2002). Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.* **2**, 0048.

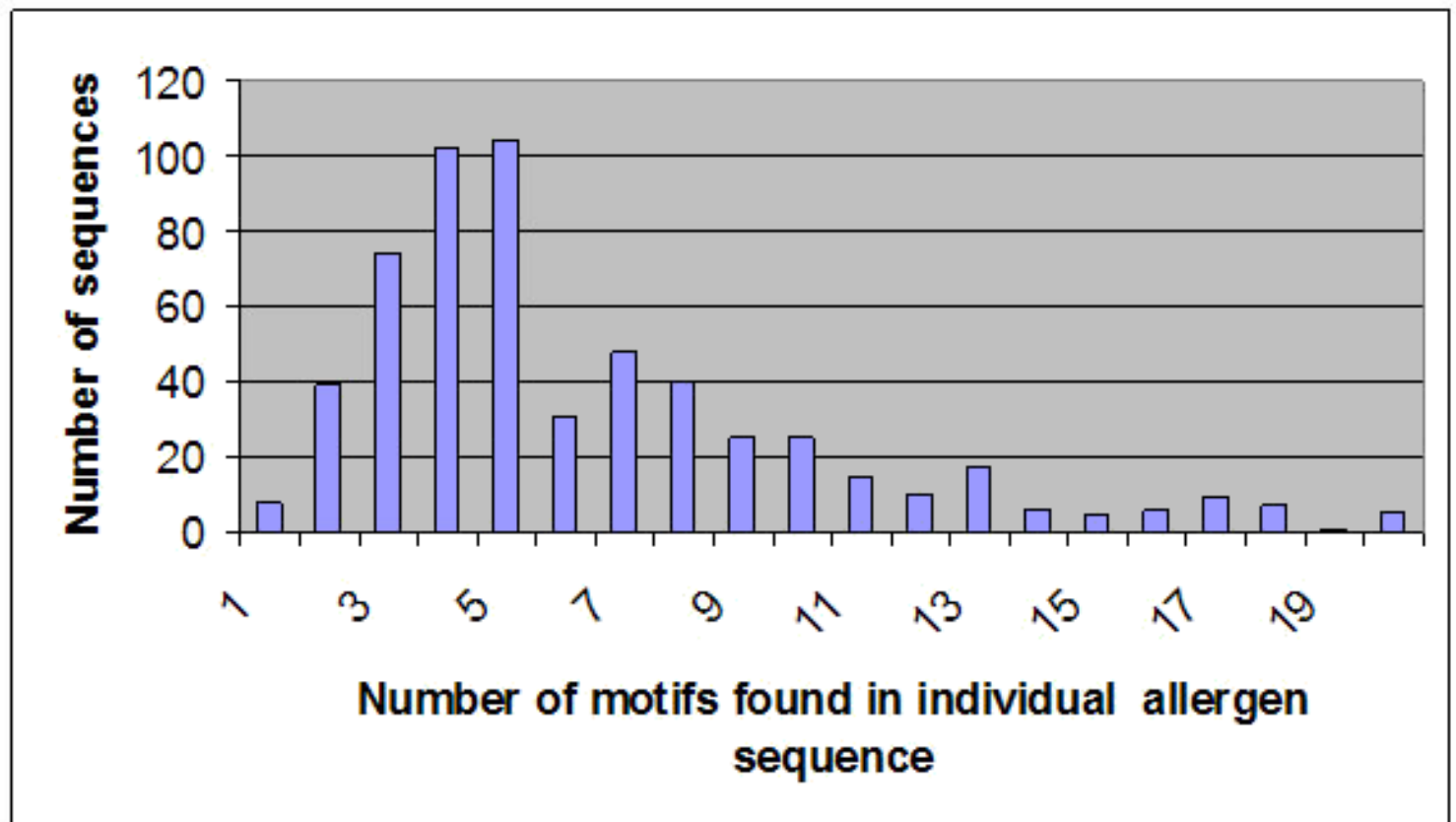


Figure 1: Histogram showing the distribution of the number of motifs extracted from the allergen sequences. 98.6% of the allergen dataset (575 sequences) contain 2 or more motifs.

Kong *et al.*, *In Silico Biol.* 7, 0006 (2006); Figure 2

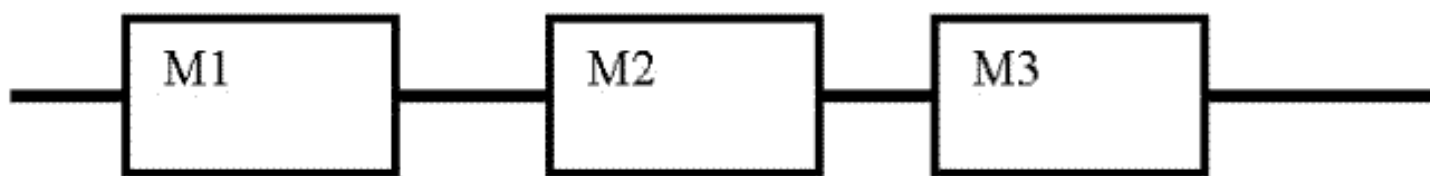


Figure 2: Motifs M1, M2, M3 found in an allergenic protein.

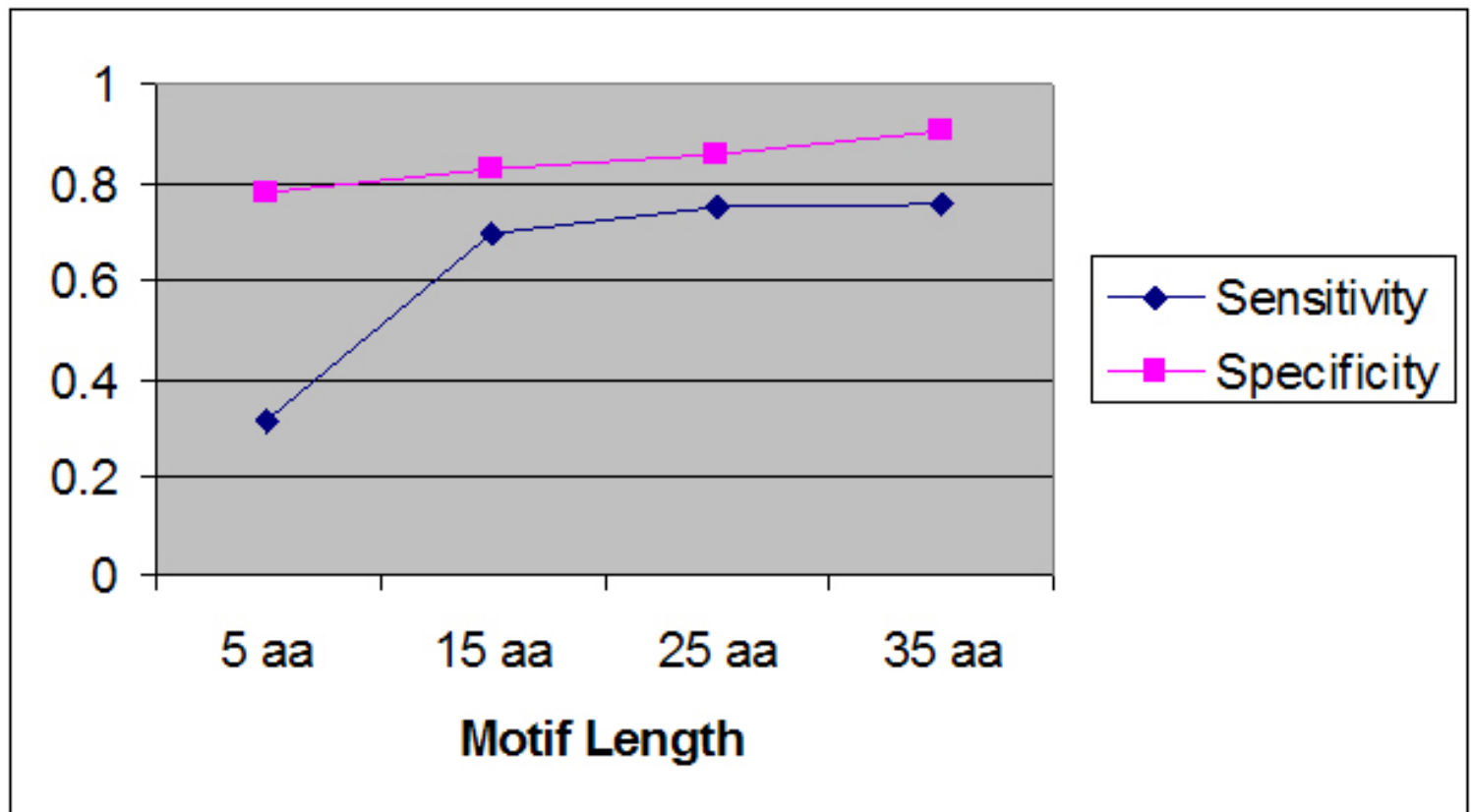


Figure 3: Plots of sensitivity, specificity vs motif length.

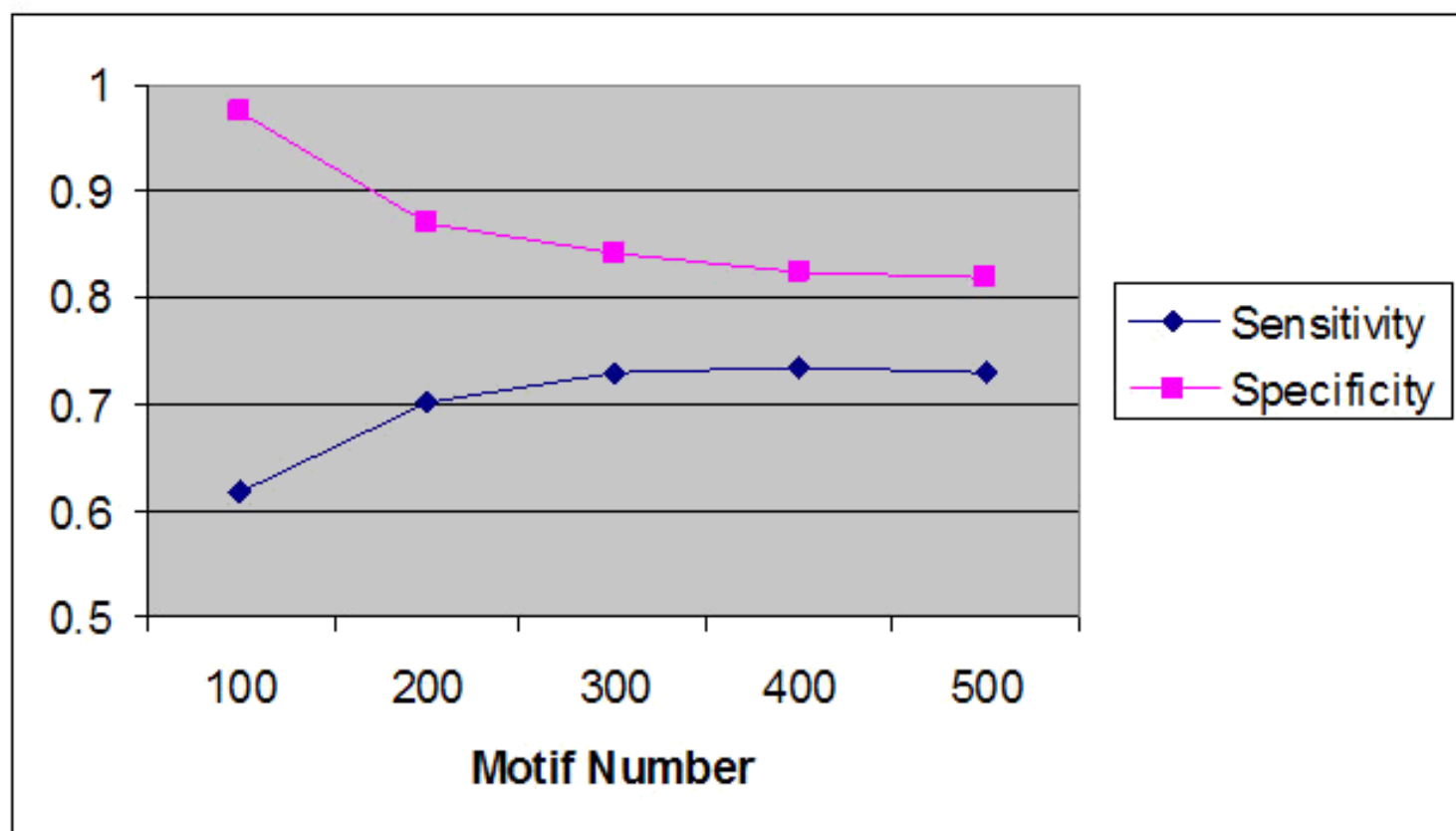


Figure 4: Plots of sensitivity, specificity vs motif number.

Kong *et al.*, *In Silico Biol.* 7, 0006 (2006); Figure 5

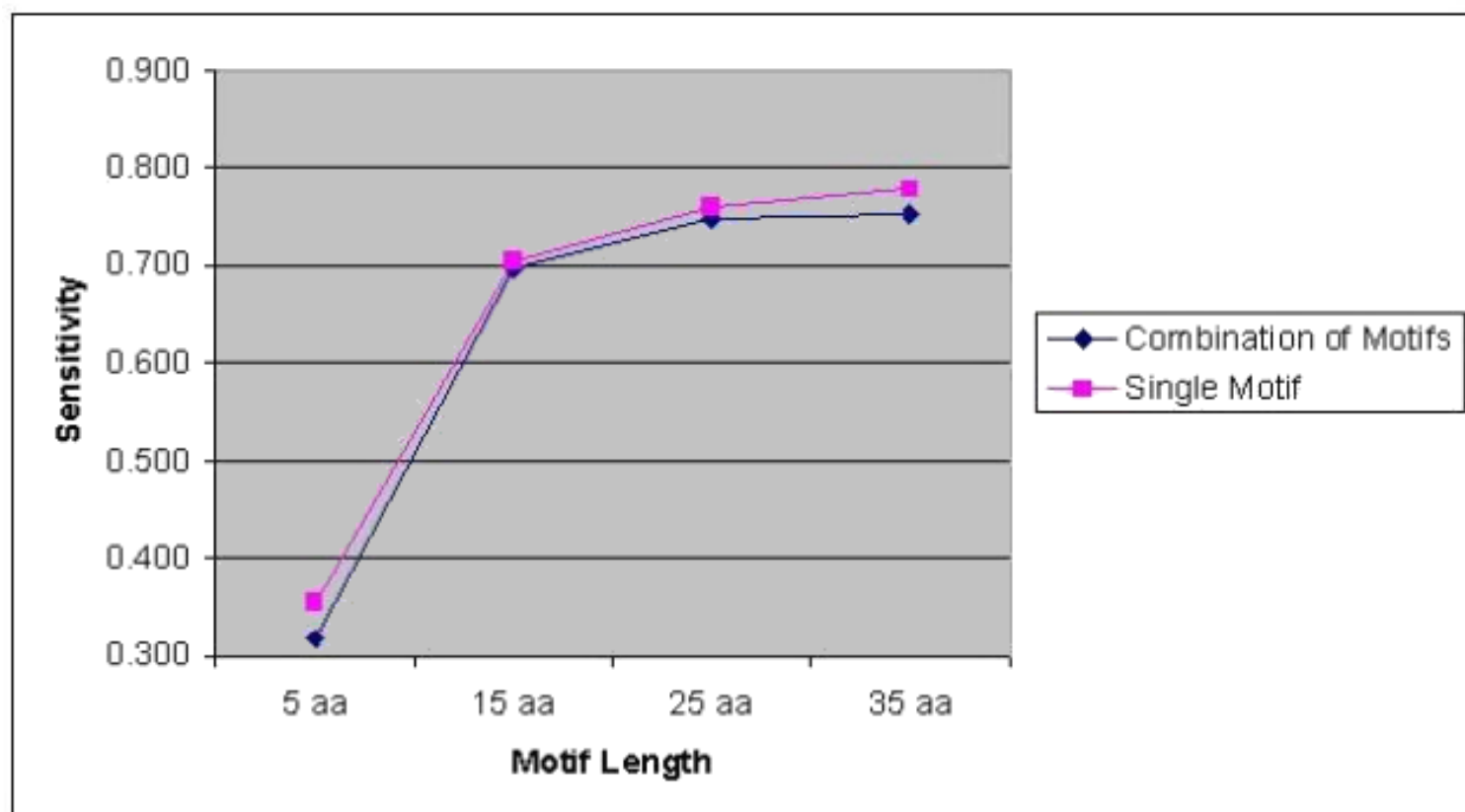


Figure 5: Plot of sensitivity vs motif length for combination of motifs and single motif algorithms.

Kong *et al.*, In Silico Biol. 7, 0006 (2006); Figure 6

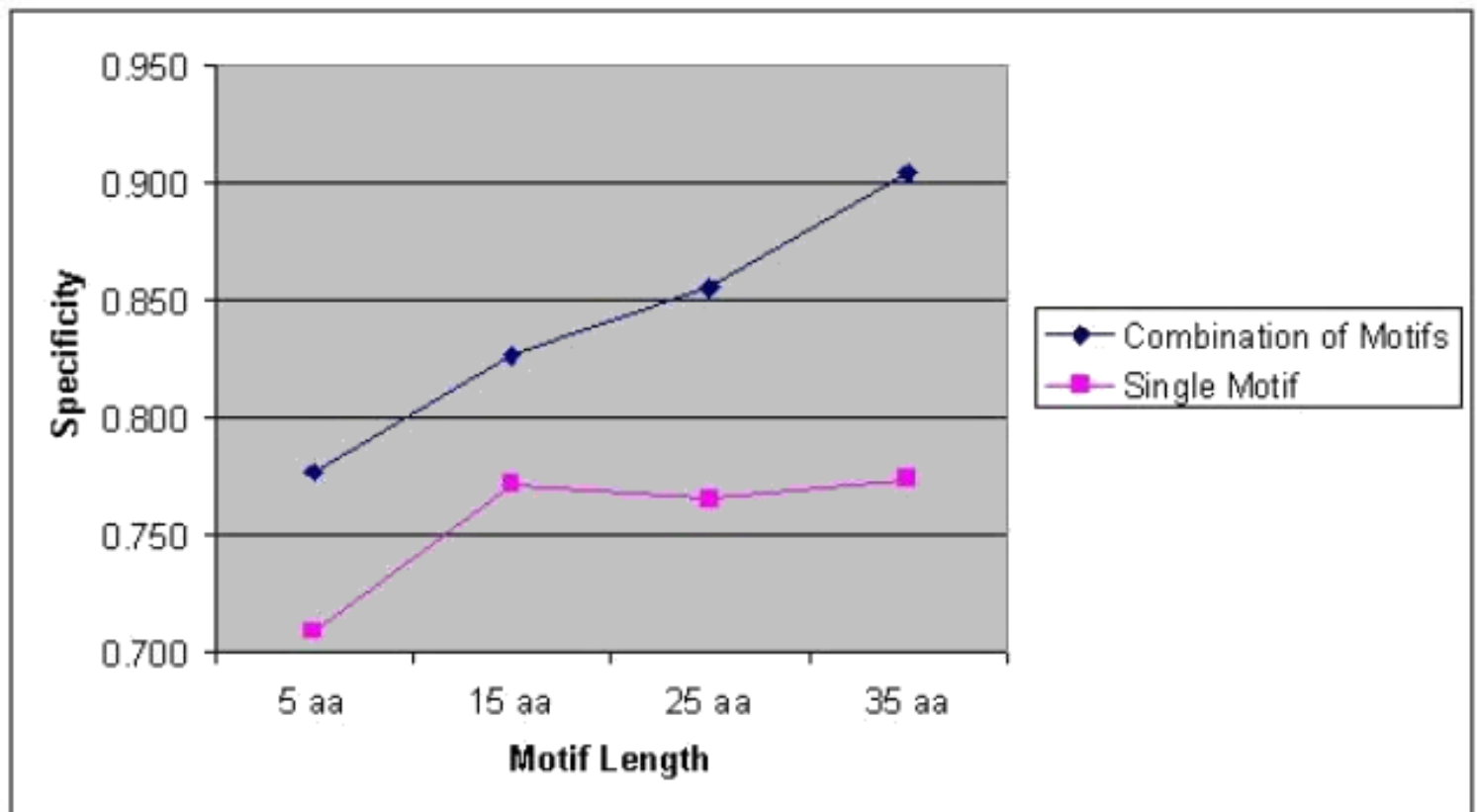


Figure 6: Plot of specificity vs motif length for combination of motifs and single motif algorithms.