Jenny Radosevich
Biochemistry 218 Final Project
June 5, 2002

Structural Considerations are Important when Performing Computer-Based Searches for Molecular Mimics

Introduction

Myelin basic protein (MBP)-reactive T cells are thought to play an important role the autoimmune disease Multiple Sclerosis (MS). MBP is an abundant protein on the myelin sheath which has a role in both myelin sheath formation and stabilization. MS patients have autoantibodies and autoreactive T cells which recognize MBP as foreign and thus attack the myelin sheath, essentially short circuiting the nervous system. One important question is: how do MS patients obtain these autoreactive T cells? In the process of negative selection, T cells which recognize self-antigens are supposed to be deleted. A leading explanation is that molecular mimicry plays a role in the induction of autoimmunity. In this hypothesis, infections with viruses or bacteria which contain similar sequences as MBP stimulate a series of T cells which cross-react with self-MBP. Either these bacterial or viral infections remain in the host, allowing for the continual activation of autoreactive T cells. Or, the chronic destruction of the myelin sheath leads to increased MBP fragmentation and exposure of antigenic epitopes which cause the chronic stimulation of autoreactive T cells. In either case, it is thought that bacteria and/or viruses induce the autoimmune state and thus their identity is desired.

An extensive study was done to determine the most important residues in the MBP (85-99) epitope and possible bacterial and viral molecular mimics (Wucherpfennig and Strominger 1995). Using amino acid substitution, the tolerated residues for MBP-reactive T-cell stimulation were identified. A motif was generated and the viral and bacterial databases of 1995 were searched. Of 129 pathogenic bacteria and viruses identified, 5 sequences were found that actually stimulated MBP-reactive T cell clones isolated from MS patients. This low ratio of true positives to total number of peptides which fit the motif displays that motif searching alone is not the best way to search for molecular mimics. Furthermore, MBP is not the only target of autoimmune activity in MS patients. There are other myelin antigens which are important such as myelin oligodendrocyte glycoprotein and proteolipid protein. More efficient methods are needed for the discovery of microbial molecular mimics of human proteins.

In this study, I first used multiple sequence alignment to derive a consensus sequence using the SeqWeb program Pretty to determine the residue specificity and sequence information contained in 5 true molecular mimics (Wucherpfennig and Strominger 1995). Knowing that the structure of the peptide also affects MHC/TCR binding, I next compared the primary and secondary structure of known true positives to other peptides which fit the authors' motifs, but which did not stimulate the autoreactive T cells. I used PepPlot and Peptide Structure Prediction from SeqWeb on the Stanford PMGM server and Jpred[2] on the internet through EBI. PHD and Swiss-Model did not work as the peptides were too short for analysis. I found that a peptide's structure was an important component of its ability to stimulate a T cell clone and proposed a method to incorporate structure when searching for new molecular mimics using motifs. I next attempted to use

position specific scoring matrices such as eMATRIX to predict molecular mimics, although more true positive sequences were necessary. Finally, I proposed a new method which may more accurately predict true positive molecular mimics. Yet, this method will only be successful if around 30 true positives are already known and if basic structural requirements are defined.

Results and Discussion

Detection of Consensus Sequences

In retrospect of Wucherpfennig and Strominger's project, I wanted to determine how much useful information was contained in the 5 true molecular mimics. First, using the PMGM SeqWeb program Pretty, I made a consensus sequence of the MBP peptide residues 85-99 and experimentally tested molecular mimics which were shown to stimulate MBP-reactive T cells. Pretty was performed using Blosum 62 and 3 minimum votes were required for a consensus (figure 1).

Interestingly, using only these 6 sequences, the most important residues were identified. The most important residues are the T cell contact residues which have been experimentally determined to be His-90, Phe-91, and Lys-93 (Wucherpfennig and Strominger 1995). Val-88 is an important T-cell receptor (TCR) contact residue in some MBP-reactive T cell clones (Wucherpfennig and Strominger 1995) and this residue is represented in half the sequences. However, I have to lower the minimum votes required for a consensus in order to have this residue contained in the consensus sequence. Another important contact residue contained in the consensus sequence is Val-89. This residue is an important MHC contact residue in both HLA-DR2 and HLA-DQ1 subtypes. Since MHC binding is much more degenerate, it is no surprise that all major TCR contact residues show up in a consensus sequence while only some MHC contact residues are represented. Therefore, knowing only a few true positive molecular mimics is enough for programs like Pretty to identify the important residues using a multiple sequence alignment. In this case, it picked out the most important TCR contact residues which are known from biological experiments. I was pleased that such a good consensus was formed from only 5 molecular mimics.

I next compared the consensus formed from these mimics with that formed from a greater number of true positives- a combination of these 5 mimics and 12 mimics found in another MBP molecular mimicry study (Hemmer *et.al.* 1997). I increased the minimal votes required for a consensus from 3 to 7 since I was more than doubling the number of sequences. Strangely, when Pretty performed the multiple alignment of these sequences (figure 2), it aligned the Influenza A virus and Herpes Simplex Virus (HSV) sequences differently than in figure 1.

To explore the sequence alignment using a different program, I used EBI ClustalW to align this same group of peptides (figure 3). I could not use ClustalW on the Decypher machine as some of my sequences were less than 15 residues. ClustalW aligned the

sequences the same way Pretty did in figure 1. Therefore, these two alignments display a difference in the Pretty pileup and ClustalW methods of multiple sequence alignment.

Since I obtained a different alignment with the larger group of mimics, it is possible that the larger group of mimics was dominating the group. The consensus sequence from this combined group contained only some, but not all of the important TCR contact residues for the Wucherpfennig and Strominger molecular mimics. The molecular mimics from Hemmer *et.al.* may have had slightly different requirements for binding to the TCR on their T cell clone. Therefore, in order to keep this study accurate, I will focus on the molecular mimics identified by Wucherpfennig and Strominger as the TCR contact residues are well defined.

Primary and Secondary Structure Prediction and Comparison

The peptides which are presented by MHC proteins to the TCR do not lie on a flat surface. Rather, the peptides fit into grooves both on the MHC and TCR. Figure 3 shows a picture of MBP peptide binding to a specific T cell clone (from Steinman *et.al.* 1995). In this case, residues F, K, and N contact the TCR and occur above the plane of the overall peptide. Likewise, residues H, F, I, and R contact the different MHC pockets and lie below the plane of the peptide in this figure. MHC contacts are more degenerate, yet the TCR also has a degree of flexibility as very similar residues can substitute for one another as in MBP molecular mimicry (Wucherpfennig and Strominger 1995, Hemmer *et.al.* 1997). A peptide must contain key contact residues to stimulate a specific T cell. But, the organization or structure of the peptide in the binding cleft may be just as important for T-cell stimulation as having the correct contact residues. Hydrophobic and hydrophilic residues play an important role in the way a peptide binds to the MHC and TCR. Further, the presence of a _-sheet or _-helix could also influence the ability of certain peptides to effectively stimulate a T cell.

I wanted to predict both the primary and secondary structures of the true positive molecular mimics with that of MBP (85-99). Unfortunately, my peptides were too short to use either PHD or Swiss-Model. However, I was able to successfully use the PMGM SeqWeb PepPlot and PeptideStructure prediction programs along with Jpred, a protein prediction program through EBI. Each of these programs has different attributes.

<u>PepPlot</u>

The PepPlot program was extremely easy to use. It took my short peptide sequences, only 11-15 residues in length, and plotted both predicted secondary structure and hydropathy. The black curve in figures 5 and 6 is the Kyte and Doolittle hydropathy measure (Kyte and Doolittle 1982). This curve is the average of a residue-specific hydrophobicity index over a window of nine residues. It only starts at residue 5 because it looks at the 4 residues before and after that spot in order to calculate hydropathy. When the line is in the upper half of the frame, it indicates a hydrophobic region, and when it is in the lower half, a hydrophilic region. The hydropathy plot is especially

interesting as the incidence of hydrophobic and hydrophilic residues may contribute to the binding of the peptide in the MHC/TCR groove.

Using PepPlot, I compared the hydropathy curve of MBP and all 5 true molecular mimic peptides: Herpes simplex virus (HSV), Adenovirus12, Epstein-Barr virus (EBV), Influenza type A virus, and *Pseudomonas aeruginosa*. It was very interesting that these 6 sequences contained similar hydropathy profiles (figure 5a,b). The black curve starts from the hydrophilic side around residue 5, peaks in the neutral-hydrophobic side at around residue 8, and returns back to the hydrophilic side by residue 10. The MBP plot is included in both figures 5a and 5b for reference. All 5 of these mimics stimulate a MBP-autoreactive T cell. Yet, *Pseudomonas aeruginosa*, Herpes simplex virus, and Adenovirus12 stimulate T cell clone Hy.1B11 while EBV and Influenza type A virus stimulate autoreactive T cell clones Hy.2E11 and Hy.1G11 from the same patient. All of these T cell clones require the same important TCR contact residues on the corresponding peptide. Yet, each clone is different in that it preferentially binds different MHC subtypes. Likewise, each subgroup of mimics contains even more similar hydropathy plots (compare figure 5a and 5b).

I next wanted to determine if the false molecular mimics, those which were identified using the motif but did not stimulate the MBP-reactive T cells, contained similar hydropathy plots as true molecular mimics. Interestingly, the false molecular mimics had very different hydropathy curves than the 5 true mimics (compare figure 6 to figures 5a and 5b). Out of the 6 false negatives which I tested on PepPlot, only *Klebsiella pneumoniae* had a somewhat similar hydropathy pattern to the true molecular mimics (compare figure 6b to figure 5b).

The similar hydropathy plots of the mimics that activate the same MBP-reactive T cell clone leads me to two conclusions. First, primary structure and level of hydrophobicity and hydrophilicity along the peptide play a large role in its ability to stimulate a given T cell clone. Second, the hydropathy plot of PepPlot could possibly be used a priori such that a researcher would eliminate many of the false molecular mimics based on structure comparisons to MBP (or another peptide of interest). This elimination would reduce the number of peptides which needed to be synthesized and experimentally tested. After this structural filter, the efficiency of identification of molecular mimics would increase greatly so that many fewer peptides would need to be tested to find most true mimics.

Although the hydropathy curve showed a clear pattern, the hydrophobic moment did not seem to be correlated with the ability of a peptide to stimulate a given T cell. The hydrophobic moment curves rise when the peptide forms either an _-helix or a _-sheet at the interface between the solvent and the interior of the molecule. In other words, it is the probability that the sequence at each position is amphiphilic, having hydrophobic residues on one side and hydrophilic residues on the other. Red denotes _-helices and blue denotes _-sheets. The presence or absence of _-helices and _-sheets may indeed be important for peptide stimulation of T cells. In fact, the antigenic region of MBP, residues 85-99, forms a _-strand in the full protein according to the PDB structure

databank (http://www.biochem.ucl.ac.uk/bsm/pdbsum/1qcl/main.html). However, a clear pattern of secondary structure was not evident in this output from PepPlot.

PeptideStructure

PeptideStructure is another structure prediction program on SeqWeb. It plots a hydrophilicity curve also, yet the display is not as clear as PepPlot and differences between the peptides are difficult to detect (see figure 7a,b,c). PeptideStructure plots secondary structure based on both the Chou-Fasman method and the Garnier-Osguthorpe-Robson method. Both MBP and the 2 true molecular mimics (figure 7a,b) have predicted _-sheets in the middle of the peptide. One false mimic, Hepatitis C virus has predicted _-sheets while the other false mimic, *Klebsiella pneumoniae*, has no _-sheets (figure 7c). All peptides except for MBP have predicted _-helices.

An interesting feature of PeptideStructure is the plot of the antigenic index (AI). AI is a measure of the probability that a region is antigenic. It is calculated by summing several weighted measures of secondary structure. This feature is most likely more useful with longer peptides. There seems to be no difference in AI among any of the peptides shown in figure 7.

Jpred^2

Jpred^2, available through EMBL-European BioInformatics Institute, can also be used to predict secondary structure. Jpred^2 predicted a helix in the middle of the peptides; however it aligned the sequences with gaps to make the prediction which is not accurate. Nonetheless, another program called PSIPRED which is accessible through the EMBL Predict Protein site also placed a helix in the middle of MBP 85-99 (figure 8).

The best part of Jpred^2 was not actually its ability to predict secondary structure. It contains a link into Jalview which has a java formatted interactive display. Within the applet window, the viewer can change the color scheme of the residues based on hydrophobicity, helix propensity, strand propensity, and more. Also, Zappo colors can be used to visualize the properties of the specific amino acids at each position in the Clustal alignment. This program was unique in that it allowed me to enter multiple sequences rather than just one to view structural properties. Furthermore, it was very useful to have the interactive interface to view patterns. The hydrophobicity/hydrophilicity pattern which I first saw in PepPlot is evident in Jalview. Unfortunately, there are some problems sending the postscript file by e-mail or saving it to disc at this time so the display only exists in an applet window. The current link to my Jpred^2 output is http://jura.ebi.ac.uk:8888//jpred-bin/chklog?0829_4530 The Jalview applet window is accessible through this link.

Other Structure Prediction Programs

I attempted to use other structure prediction programs such as Swiss-model and Predict Protein PHD. However, the peptide sequences were too short for analysis. Peptides need to be at least 25 residues long for most structural programs. Peptides longer than 20 amino acids are rarely presented to T cells and so these prediction programs requiring more residues are useless for this type of study. Nonetheless, the Predict Protein website was useful as it contained links to multiple protein structure sites and automatically sent my query to other structural programs.

Combination of Motif Scans and Structure Prediction

I think that a program which a ran motif scan and allowed you to select certain sequences and view their primary and secondary structure predictions would be very helpful in eliminating sequences which fit a motif, but do not have the correct structure or types of flanking residues (hydrophobic or hydrophilic) to stimulate a particular T cell. Biologically testing hundreds of peptides to look for molecular mimics is labor intensive and thus these studies are rare. Furthermore, only around 4-8% of the predicted molecular mimics identified from the motif actually stimulate a given T cell clone (Wucherpfennig and Strominger 1995, Grogan *et.al.* 1999). I think that comparing the structure of these peptides computationally before synthesizing and testing the peptides in vitro would be a huge time saver and increase the efficiency of searching for mimics. Graphically comparing each sequence at least at the level of hydrophobicity could really decrease the amount of peptides one needs to biologically test. Yet, comparing each sequence one by one is time consuming in itself.

It would be helpful to somehow input the results from the motif search directly into a program like ProPlot so as to compare them easily. The Decypher machine allows the input of one group of sequences from one program into another program. An interface that could bridge the motif database search with a structural program containing the ability to overlay many peptides on the same graph would be ideal. The researcher could then observe patterns of hydrophobicity and/or areas of the peptide containing helices or sheets to determine which peptides have the same structure as the peptide target of an autoreactive T cell.

Position Specific Scoring Matrices

Another way to increase the efficiency of motif-based molecular mimic searches would be to generate a position specific scoring matrix (Kirk Jensen, Final Project Fall 2001). Yet, one needs to begin with a well-defined system. Grogan *et.al*. tested 832 peptides on the same autoreactive mouse T cell clone and found 61 true molecular mimics which cross-reacted with an autoreactive T cell clone. Kirk used these mimics to show that eMATRIX is a more efficient way than using a motif pattern to identify new molecular mimics. I have too few true positives to generate an accurate position specific scoring

matrix. Yet, if I had a few more true positive molecular mimics of the same T cell clone, I could generate an eMATRIX and possibly more accurately search for new molecular mimics.

Proposal of an Accurate Prediction Tool <u>if</u> Structure and True Positives are Defined

eMATRIX generates a weighted matrix based on the amino acid frequencies in the training set. As explained in this paper, both structural information and functional information such as TCR contact residues are critical to a true molecular mimic motif. Therefore, it may be useful to alter the eMATRIX program as T cell contact residues are very important and deserve more weight, thus generating a more accurate position-specific scoring matrix for molecular mimicry prediction. To fulfill this idea, I could generate a molecular properties matrix. This matrix would be based on contact residues or other structural properties and could weight eMATRIX if I multiplied this matrix by the normal eMATRIX generated by true positive molecular mimics. Therefore, the resulting matrix would be weighted by both amino acid frequency and structural properties. If I had thought of this project earlier, I may have used the Grogan *et.al.* dataset to test the efficiency of regular eMATRIX versus weighted (multiplied) eMATRIX in accurately predicting molecular mimics, thus extending Kirk's project from Fall 2001. Nonetheless, my present study demonstrated the importance of structure in finding true molecular mimics which led to the idea that a combination of structural information along with motifs or position specific scoring matrices would be a more accurate way of predicting molecular mimics.

Jenny Radosevich
June 5, 2002

Figure 1. Pretty Multiple Sequence Alignment and Consensus Sequence of MBP and Five Molecular Mimicry Sequences.

```
             1               15
          HSV FRQLVHFVRD FAQLL
InfluenzaAvi YRNLVWFIKK NTRYP
Adenovirus12 DFEVVTFLKD VLPEF
           P DRLLMLFAKD VVSRN
85-99wholeMB ENPVVHFFKN IVTPR
          EBV TGGVYHFVKK HVHES
    Consensus -R--VHF-KD -V---
```

Figure 2. Pretty Multiple Sequence Alignment and Consensus Sequence of MBP and Molecular Mimicry Sequences from Two Different Studies.

```
             1                       22
Adenovirus12 DFEVVTFLKD VLPEF~~~~~ ~~
          f2 ~~~DILILKL VVGE~~~~~~ ~~
           P ~DRLLMLFAK DVVSRN~~~~ ~~
          s3 ~~~VAMLMKN TIIA~~~~~~ ~~
          h1 ~~~DLIFYRN VV..IK~~~~ ~~
          h2 ~~~DLIFYKN VV..IK~~~~ ~~
          h3 ~~~DLIMYKN VV..IK~~~~ ~~
          h4 ~~~DLIMYRN VV..IK~~~~ ~~
          h5 ~~~DLIMYRN VV..IA~~~~ ~~
InfluenzaAvi ~~~~~~~YRN LVWFIKKNTR YP
          s1 ~~~QVNQFKN VIFE~~~~~~ ~~
          s2 ~~~AVKGFRN VIIG~~~~~~ ~~
85-99wholeMB ENPVVHFFKN IVTPR~~~~~ ~~
          f1 ~~~WRKFFKN VVSS~~~~~~ ~~
          f3 ~~~AGSFFKN PVVA~~~~~~ ~~
          s4 ~~~WIHQLKN VIRY~~~~~~ ~~
         HSV ~~~~~~~FRQ LVHFVRDFAQ LL
         EBV TGGVYHFVKK HVHES~~~~~ ~~
   Consensus ------FFKN VV-------- --
```

Figure 3. EBI ClustalW Multiple Sequence Alignment of MBP and Molecular Mimicry Sequences from Two Different Studies.

```
CLUSTAL W (1.82) multiple sequence alignment


p.aeruginosa      DRLLMLFAKDVVSRN 15
f2                D---ILILKLVVGE- 11
adenovirus        DFEVVTFLKDVLPEF 15
mbp               ENPVVHFFKNIVTPR 15
f3                ---AGSFFKNPVVA- 11
f1                ---WRKFFKNVVSS- 11
s1                ---QVNQFKNVIFE- 11
s2                ---AVKGFRNVIIG- 11
s4                ---WIHQLKNVIRY- 11
h4                ---DLIMYRNVVIK- 11
h5                ---DLIMYRNVVIA- 11
h3                ---DLIMYKNVVIK- 11
h1                ---DLIFYRNVVIK- 11
h2                ---DLIFYKNVVIK- 11
influenza         YRNLVWFIKKNTRYP 15
s3                ---VAMLMKNTIIA- 11
ebv               TGGVYHFVKKHVHES 15
hsv               FRQLVHFVRDFAQLL 15
```

Figure 4. The spatial orientation of MBP residues 84-103 in the MHC-TCR binding groove. (From Steinman *et.al.* 1995).

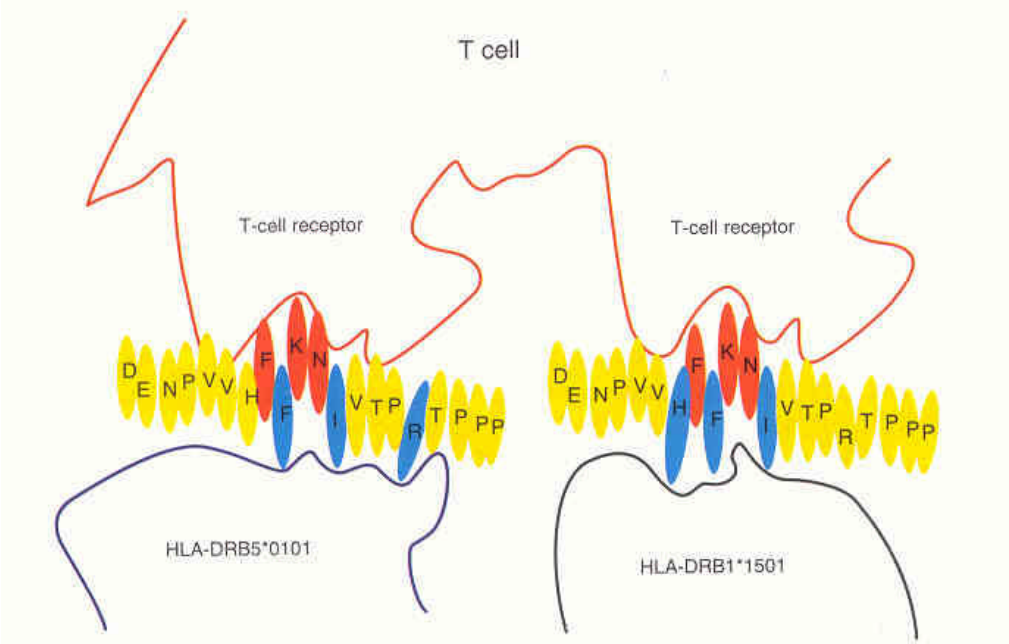


Figure 1. MHC and TCR contact sites for myelin basic protein peptide p84–103. The key MHC (blue) and T-cell receptor (red) contacts are shown for two HLA-DR2 subtypes associated with MS.

Jenny Radosevich
June 5, 2002

Figure 5a. PepPlot hydrophobic moment and hydropathy curves of MBP and three **true molecular mimics** which activate the same autoreactive T cell clone, Hy.1B11.

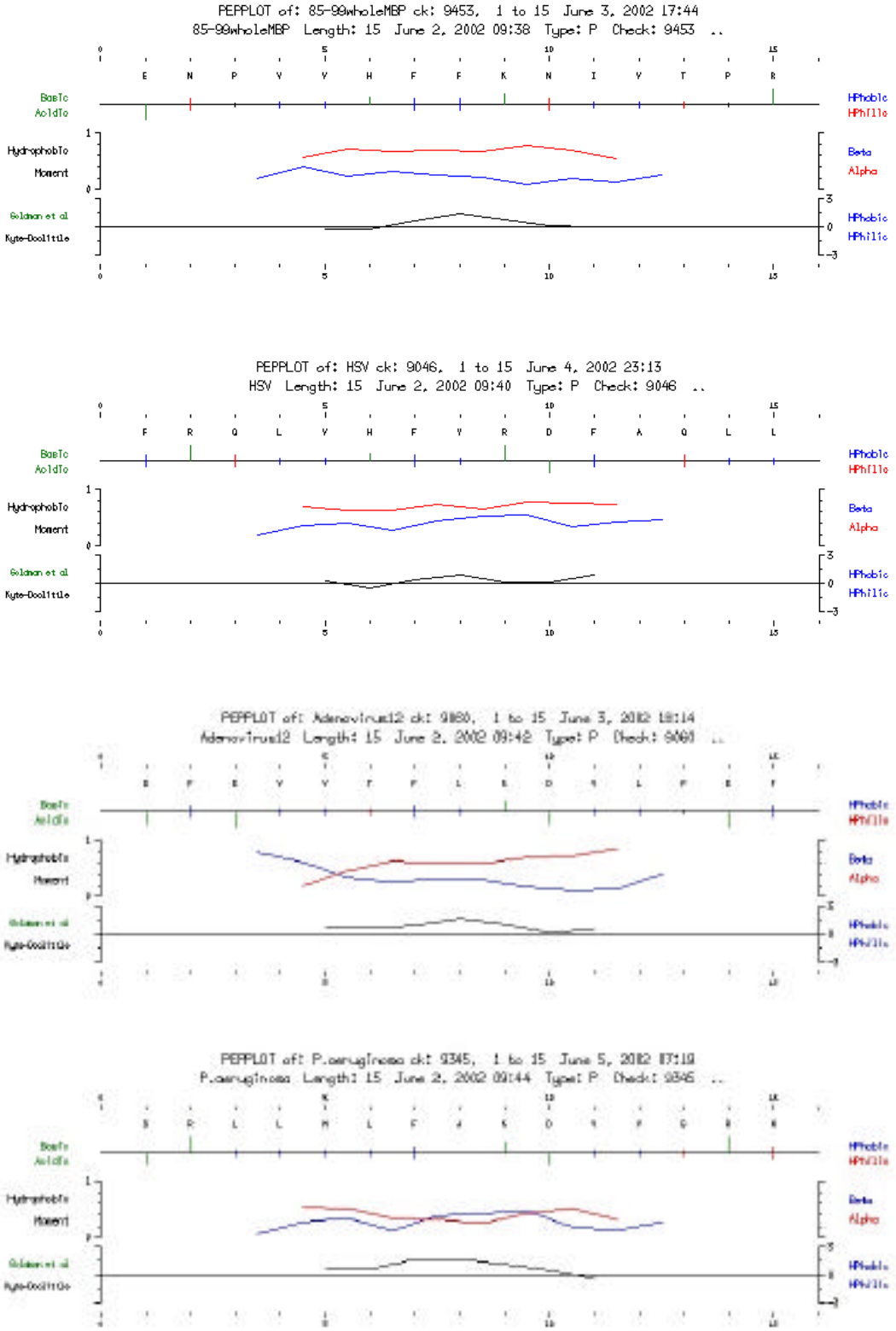


Figure 5b. PepPlot hydrophobic moment and hydropathy curves of MBP and two **true molecular mimics** which activate the same autoreactive T cell clones, Hy.2E11 and Hy.1G11.

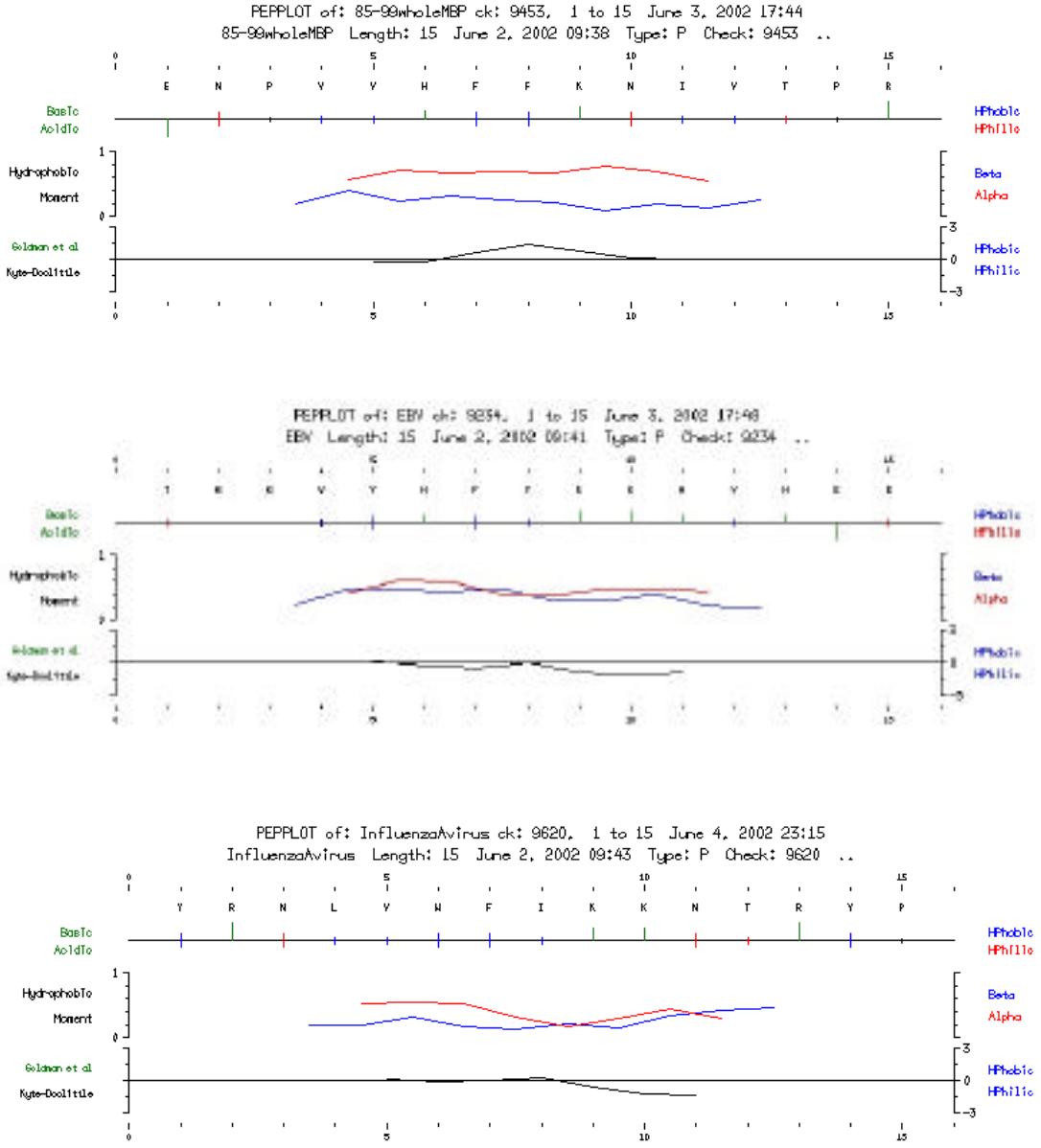


Figure 6a. PepPlot hydrophobic moment and hydropathy curves of **false molecular mimics** which were predicted by the motif and biologically tested. They did not activate an autoreactive T cell clone.

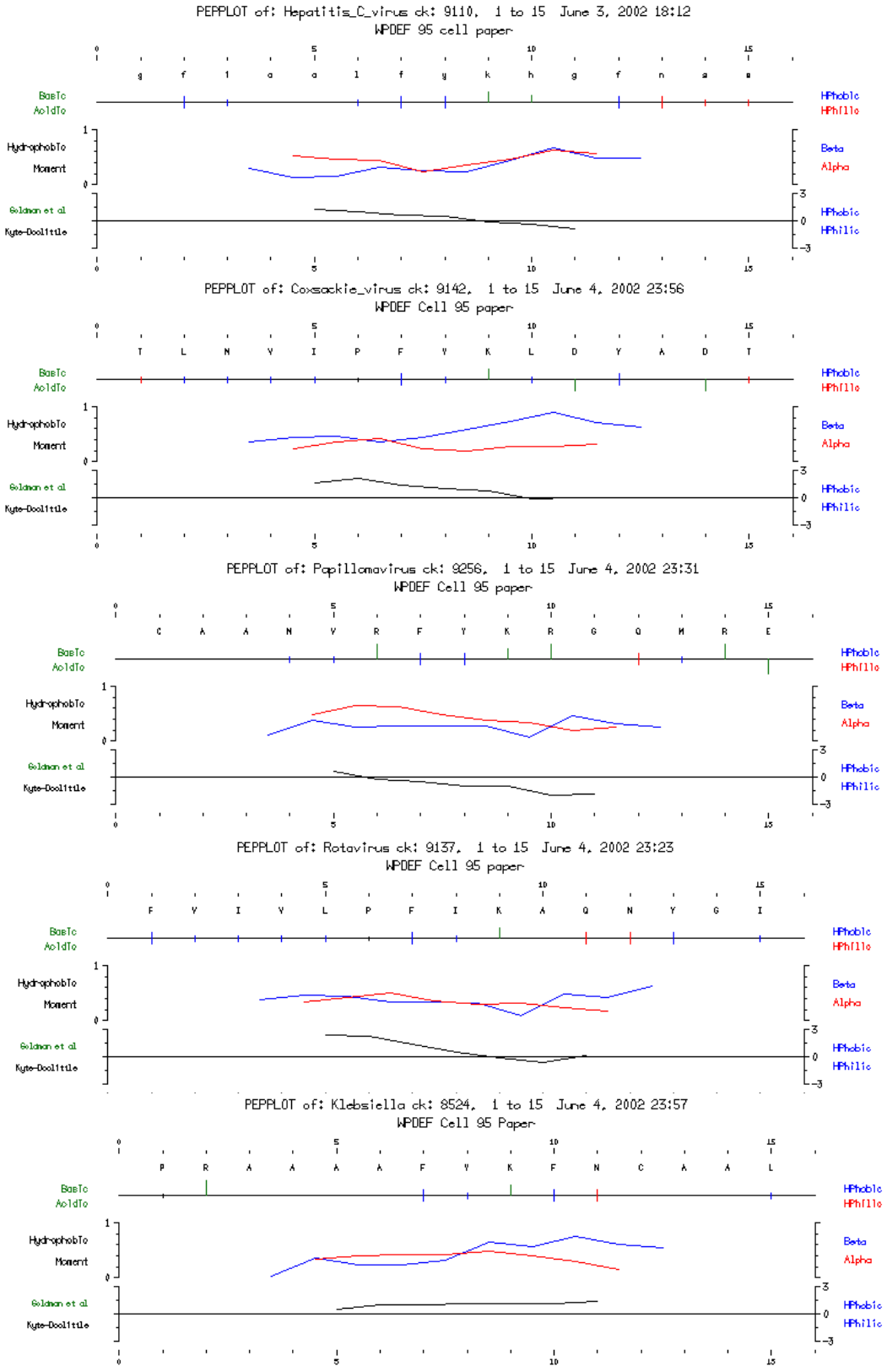


Figure 6b. *Klebsiella pneumoniae* contains a PepPlot hydropathy curve somewhat similar to true molecular mimics. Yet, it did not activate an MBP-autoreactive T cell clone.

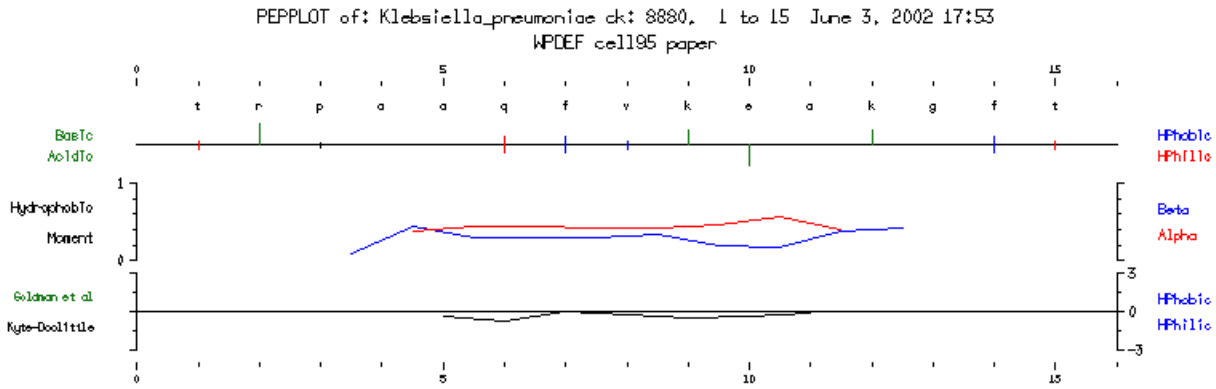


Figure 7a. Peptide Structure output for **MBP** 85-99.

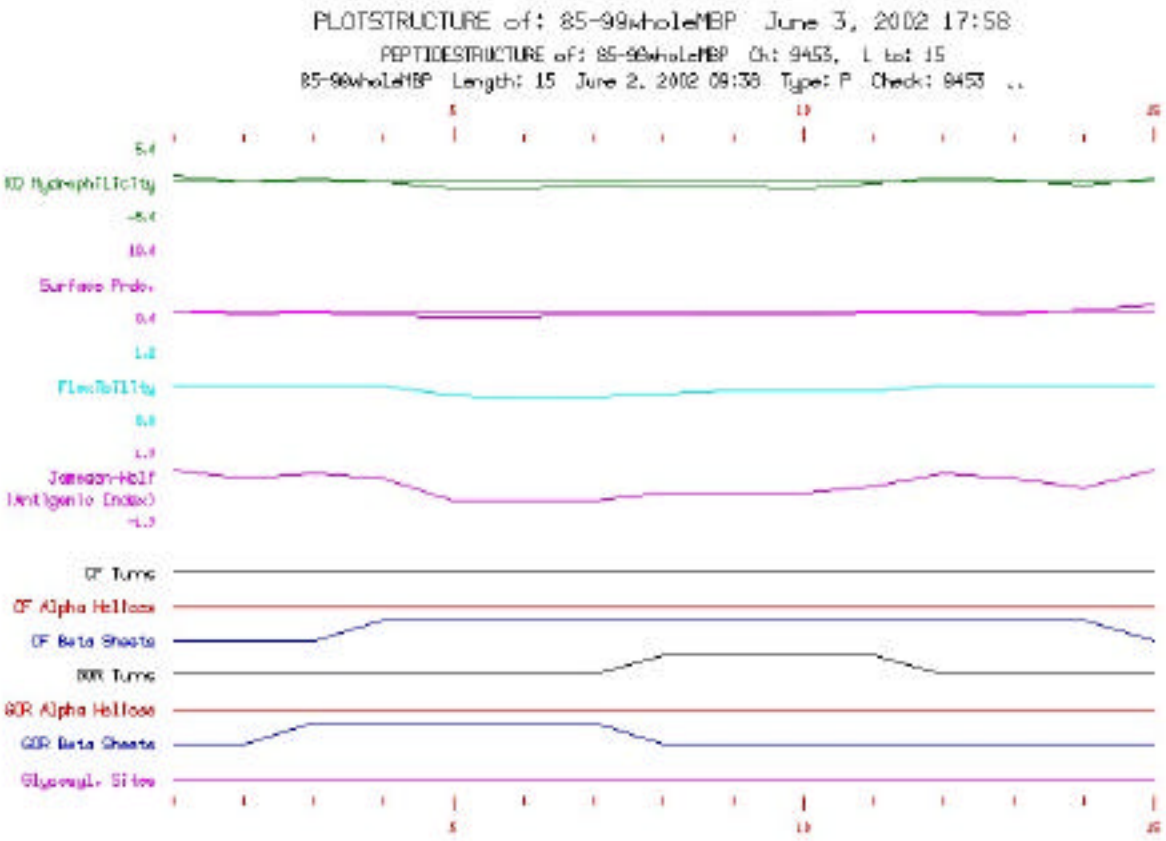


Figure 7b. Peptide Structure output for **true** molecular mimics: Adenovirus12 and EBV.

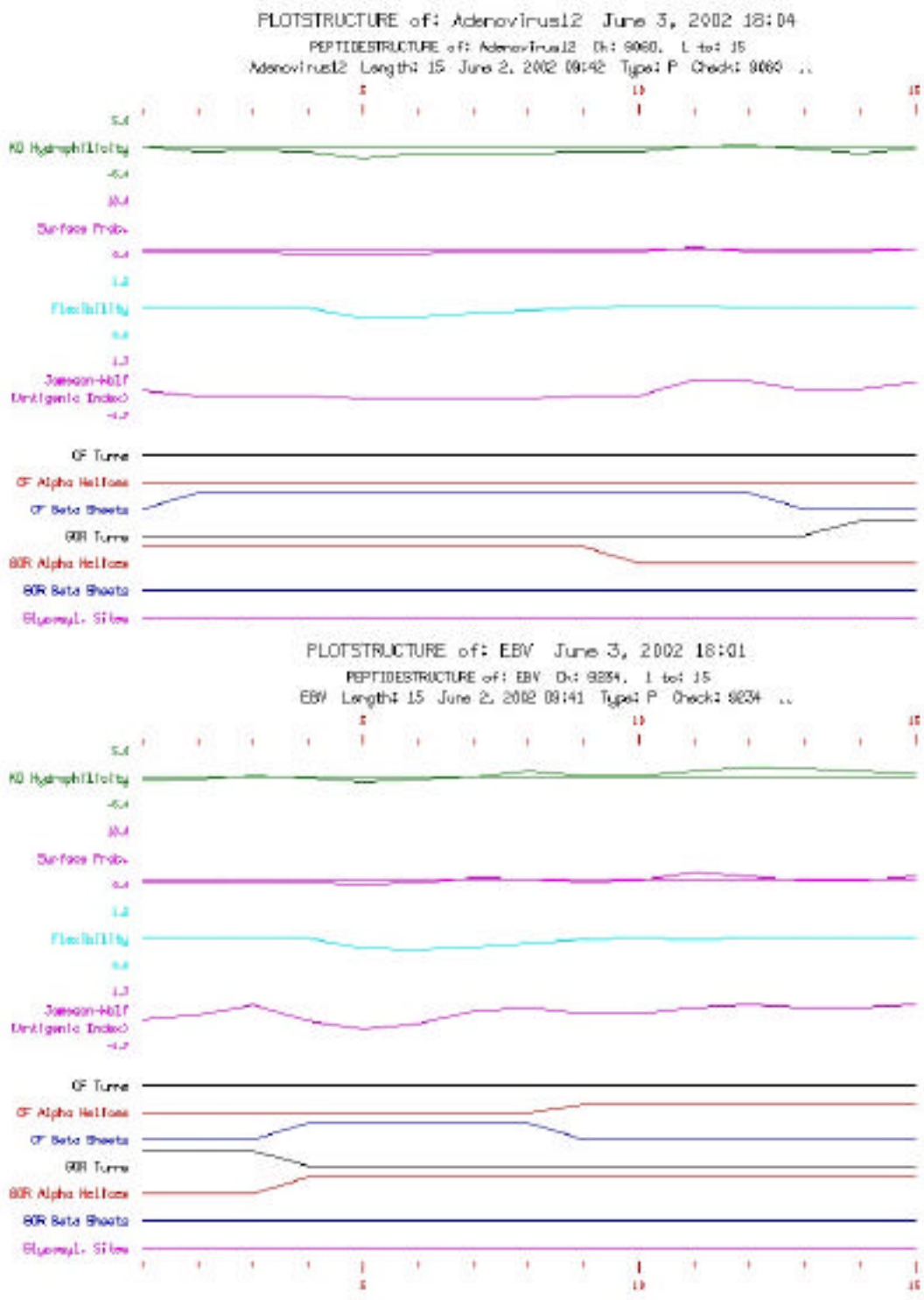


Figure 7c. Peptide Structure output for **false** molecular mimics: *Klebsiella pneumoniae* and Hepatitis C virus.

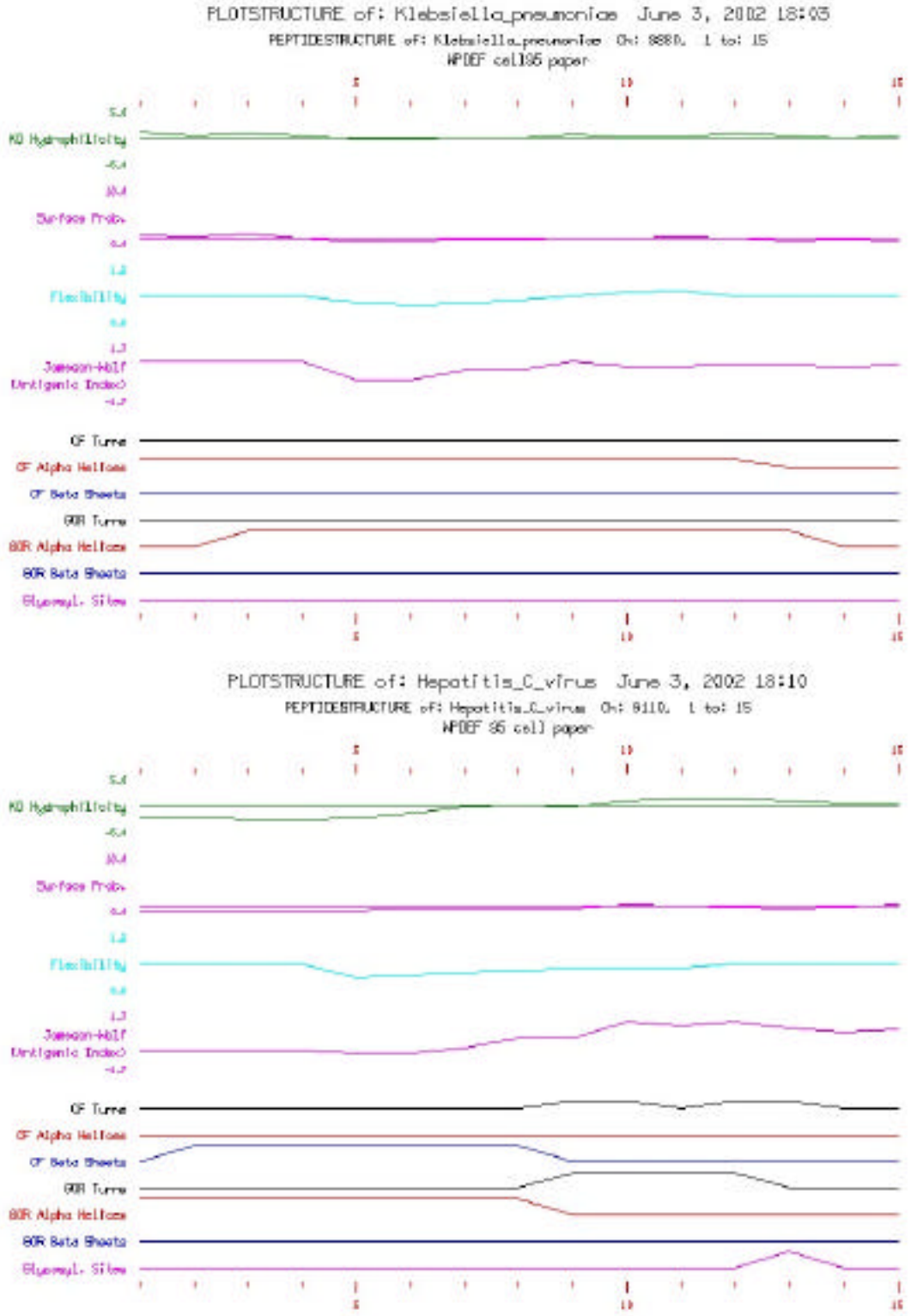
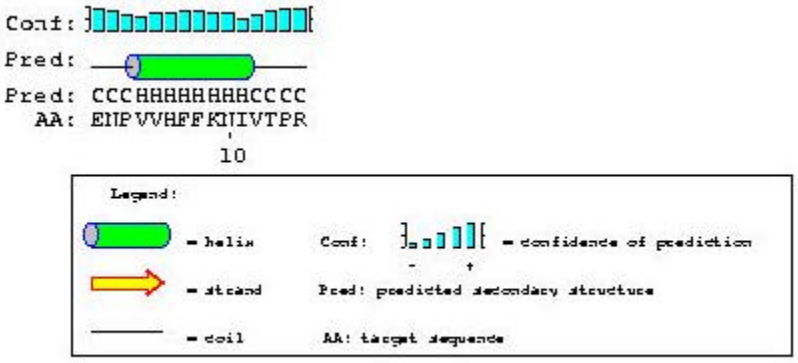


Figure 8. PSIPRED Secondary Structure Prediction Results for MBP 85-99.



References

Grogan JL, Kramer A, Nogai A, Dong L, Ohde M, Schneider-Mergener J, Kamradt T. Cross-reactivity of myelin basic protein-specific T cells with multiple microbial peptides: experimental autoimmune encephalomyelitis induction in TCR transgenic mice. J Immunol. 1999. Oct 1;163(7):3764-70.

Hemmer B, Fleckenstein BT, Vergelli M, Jung G, McFarland H, Martin R, Wiesmuller KH. Identification of high potency microbial and self ligands for a human autoreactive class II-restricted T cell clone. J Exp Med. 1997. May 5;185(9):1651-9.

Jensen K. Using Minimal-Risk Scoring Matrices to Search for "Molecular Mimics" that Can Stimulate Murine Autoreactive T cells Against MBP Ac1-11. Computational Molecular Biology Final Project. Fall 2001.

Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982. May 5;157(1):105-32.

Steinman L, Waisman A, Altmann D. Major T-cell responses in multiple sclerosis. Molecular Medicine Today. 1995. 1:79-83.

Wucherpfennig KW, Strominger JL. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. Cell. 1995. Mar 10;80(5):695-705.