Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W315–W319***
doi:10.1093/nar/gki374

MAVL/StickWRLD for protein: visualizing protein sequence families to detect non-consensus features

William C. Ray*

Children's Research Institute and The Department of Pediatrics, The Ohio State University, 700 Children's Drive, Columbus, OH 43205, USA

Received February 11, 2005; Revised and Accepted March 7, 2005

ABSTRACT

A fundamental problem with applying Consensus, Weight-Matrix or hidden Markov models as search tools for biosequences is that there is no way to know, from the model, if the modeled sequences display any dependencies between positional identities. In some instances, these dependencies are crucial in correctly accepting or rejecting other sequences as members of the family. MAVL (multiple alignment variation linker) and StickWRLD provide a web-based method to visually survey the model-training sequences to discover and characterize possible dependencies. Initially introduced for nucleic acid sequences, with MAVL/StickWRLD, it is easy to distinguish typical DNA or RNA structural dependencies in input families, identify mixed populations of distinct subfamilies, or discover novel dependencies that result from binding interactions or other selective pressures [W. Ray (2004) *Nucleic Acids Res.*, 32, W59–W63]. Since the announcement of MAVL/StickWRLD for nucleic acids, one of the most requested new features has been the extension of this visualization method to support protein alignments. We are pleased to report that this extension has been successful, that the basic visualization has been augmented in several ways to enhance protein viewing, and that the results with protein alignments are even more dramatic than with NA alignments. MAVL/StickWRLD can be accessed at http://www.microbial-pathogenesis.org/stickwrld/.

INTRODUCTION

Compact models that convey the important characteristics of a known family of biosequences are a fundamental need in the biological/life sciences. Such models are a necessity both for the purpose of describing the family, and to provide a way of conveying the required characteristics to a search algorithm so that statistically similar and possibly homologous sequences can be discovered. The models in most common use are sequence consenses (1), positional weight (probability or log-odds) matrices (1) and hidden Markov models (HMMs) (2). Through these, and variants on their basic technology, relatively sophisticated statistics can be brought to bear on the positional identities in a sequence family. Unfortunately, all suffer from the same basic statistical defect in that they all treat each position in an alignment as linearly separable in identity from every other position in the alignment. This is clearly not a valid assumption for molecules such as structural RNAs, where strict base-pairing requirements that result from structure, make the base identity at certain positions completely dependent on the bases found at other positions. While the situation with proteins is less obvious, seminal binomial mutagenesis experiments demonstrated, based on functional assays, that amino acid positional identities are not completely separable (3), and algorithms that may be able to predict compensating mutations remain an active area of research [(4,5) among many others].

Such statistically linked positions, when recognized, are typically dealt with in an ad hoc fashion by either developing special scoring techniques that treat different sections of an alignment using separate algorithms, or by excluding half of each pair of positions that are clearly not separable from the scored pattern (6,7).

Despite the statistically flawed results that occur when families containing interpositional dependencies are treated as having linearly separable positional identities, there remains no convenient way for researchers to survey alignments to discover whether such dependencies appear to exist in their sequences. MAVL and StickWRLD were developed to provide a rapid visual survey and exploration system for investigating possible interpositional dependencies. Originally developed to visualize nucleic acid (NA) alignments, MAVL and StickWRLD have been updated to enable protein alignment visualization as well. Submission of a protein (or NA) alignment to the MAVL/StickWRLD webserver produces

*Tel: +1 614 355 3522; Fax: +1 614 722 2818; Email: ray@biosci.ohio-state.edu

a three-dimensional VRML (Virtual Reality Modeling Language, http://www.web3d.org/x3d/specifications/vrml/) graph of the alignment, showing all possible amino acid (or NA) identities at every position, and links between each occupied identity and any identities at other positions to which it appears to be statistically related. The graph itself is completely navigable in the user's VRML browser, allowing the close inspection of individual positions, or the global viewing of the patterns produced by the complete alignment. Several user-configurable parameters control the stringency of the calculated statistics, as well as provide means by which the connection between statistical relationships and possible physical property compensating mutations may be explored.

ALGORITHM

MAVL/StickWRLD for protein produces its basic estimation of interpositional relationship in the same manner as MAVL/StickWRLD for NA (8). It calculates the expected number of sequences that should share a pair of residues at a particular pair of positions, based on a positional probability matrix derived from the input sequence alignment. It then subtracts this expected value from the observed number of sequences with that characteristic to obtain a residual, for every possible pair of positions and identities in the alignment.

The significance (α, the probability of finding a given number of outliers in a random population) for each residual is calculated, and a VRML graph constructed by arraying the alignment/positional probability matrix around the surface of a cylinder, and connecting those positional pairs whose residuals exceed either specific absolute (user configurable) values, or that have significance better than a (user configurable) cutoff. The positional probability of a member position in the matrix is depicted in the VRML world as a sphere with its diameter scaled to the related probability. The axial ordering of the residues is based on one of four user-selectable physical parameters; Grantham residue volume (9), Grantham residue composition (9), Grantham polarity (9) and Kyte–Doolittle residue hydropathy (10), and can be controlled while setting up the parameters for StickWRLD model generation. Numeric parameters for the physical properties are extracted from AAIndex (11). Identities are indicated both by color and label, with the user able to select between four different coloring schemes: Branden and Tooze's classic hydrophobic-property coloring (12), RasMol's Amino and Shapely color schemes [(13), http://www.openrasmol.org/doc/rasmol.html], and the CLUSTAL X default alignment coloring scheme (14); Figure 1 displays a table of the property values and colors that are applied when different options are selected.

Links between positions are depicted as cylinders connecting them, with the diameter scaled to the residual over-, or underpopulation that was calculated for the position pair. With the NA version, color was used to indicate whether the residual was positive or negative, but with the protein version we have found color to be useful for visualizing other parameters. Therefore, with MAVL/StickWRLD for protein, solid-color links indicate positive residuals (an overpopulation of sequences share these identities), and dashed links indicate negative residuals (an underpopulation of sequences share these identities).

AA	GV	GC	GP	KDH	BH	RA	RS	CX
W	170	0.13	5.4	-0.9				
Y	136	0.2	6.2	-1.3				
F	132	0	5.2	2.8				
R	124	0.65	10.5	-4.5				
K	119	0.33	11.3	-3.9				
L	111	0	4.9	3.8				
I	111	0	5.2	4.5				
M	105	0	5.7	1.9				
H	96	0.58	10.4	-3.2				
Q	85	0.89	10.5	-3.5				
V	84	0	5.9	4.2				
E	83	0.92	12.3	-3.5				
T	61	0.71	8.6	-0.7				
N	56	1.33	11.6	-3.5				
C	55	2.75	5.5	2.5				
D	54	1.38	13	-3.5				
P	32.5	0.39	8	-1.6				
S	32	1.42	9.2	-0.8				
A	31	0	8.1	1.8				
G	3	0.74	9	-0.4				
b								

Figure 1. The residue properties and colors that are available, and applied by the 'Sort Residues By', and 'Color Residues By' interface options. The AA column lists amino acid residues by their single-letter code, and 'b' for StickWRLD's gap representation. The GV, GC and GP columns list Grantham (9) volume, composition and polarity values respectively. The KDH column lists Kyte–Doolittle (10) hydropathy. The BH column displays classic Branden–Tooze hydrophobicity-associated colors. The RA and RS columns show RasMol 'Amino' and 'Shapely' color schemes, and the CX column shows the CLUSTAL X default color scheme. Please remember that computer, monitor and VRML browser settings will all affect the actual displayed colors.

The color of the links is now keyed to the similarity of the residues that the link connects, based on the same physical parameters selected for visual residue ordering. The link scales from red, for identical residues, to blue, for residues that have maximally dissimilar properties on the selected scale (because there is no utility to a visual pause at the midpoint of this scale, we do not use the often-seen spectrum that shades from red through white to blue, but rather blend from red to blue, passing through purple in between).

ACCESS

MAVL/StickWRLD is available in a Nucleic Acid version and a Protein version. The desired interface can be selected from the top page at http://www.microbial-pathogenesis.org/stickwrld/. Sequences must be pre-aligned, and submitted in raw format to MAVL, with one sequence per line, spaces, periods or hyphens indicating gaps. There is no algorithmic restriction on the length or number of submitted sequences. However, practical considerations with respect to the visualization, as well as server resources, limit useful submissions to the neighborhood

of 200 amino acids or base pairs aligned length. Submissions containing more than ~300 individual sequences can also exceed server resources. The VRML for the StickWRLD graph is returned directly to the user's web browser, and requires that the user have one of the freely available VRML viewers installed to view it. The Cosmo (http://www.cosmosoftware.com/), Cortona (http://www.parallelgraphics.com/) and FreeWRL (http://freewrl.sourceforge.net/) viewers have been tested. Each has minor differences in how they render and navigate VRML files, and the user is encouraged to test each to find one that best suits his or her personal needs. Mac users should be aware that Cortona does not currently handle text properly on Mac OS X, so positional labels are invisible, and FreeWRL requires the file to be downloaded, saved to disk and opened from the FreeWRL application, rather than operating as a browser component.

Sample files for both protein and RNA MAVL/StickWRLD analyses, including the input alignment for the domain analyzed in this paper, are available from http://www.microbial-pathogenesis.org/stickwrld/tutorial/sticktut2.html. A video-based tutorial that explains the analysis and visualization options is available from this location as well.

RESULTS AND DISCUSSION

One of the immediately apparent differences between MAVL/StickWRLD visualizations for nucleic acid and for protein, is the density and significance of links that are found. RNA structures, for example, are primarily composed of patterns of neat stems, loops and bulges, with any given base imposing few identity restrictions other than on an opposing base to which it is paired. This simplicity of structure (at the level of position-to-position interactions) is readily apparent in the StickWRLD diagram for structural RNAs, and the pattern of large-residual links is often directly reflective of the underlying RNA structure.

StickWRLD diagrams of proteins however, immediately highlight the additional complexity of amino acid bonding and packing patterns, as well as the additional complexity of the protein alphabet as compared to the DNA/RNA alphabet. Because of this, MAVL/StickWRLD visualizations have shown their greatest utility when examining small 'signature' motifs and domains, such as those that comprise the 'domain' elements of the Pfam database (15). Figure 2, for example, shows a StickWRLD diagram of the relationships found by MAVL in the Pfam ADK_lid domain (accession no. PF05191, full 299-member alignment). The active site lid from adenylate kinase, ADK_lid, is interesting in that it is structurally related to zinc-finger domains, which are typically indicative of DNA binding motifs. This would not be readily apparent in the ADK_lid structure or alignment, except that there is a divergence in the mechanism by which Gram-positive bacteria maintain the ADK_lid structure, as compared to that employed by Gram-negative bacteria and eukaryotes. In Gram-positive bacteria, a network of hydrogen bonds that restrict the conformation of a pair of structural loops, are replaced by a tetrahedrally coordinated Zn^{2+} atom, ligated by four cysteine residues. When considered with cysteines and bound zinc, it is clear that the fold topology and coordination geometry is closely related to that of the TFIIS zinc finger domain (16).

Interestingly, despite this clear and well-published divergence between ADK_lid subfamilies, HMMs typically used to describe this motif (such as those generated by the Pfam

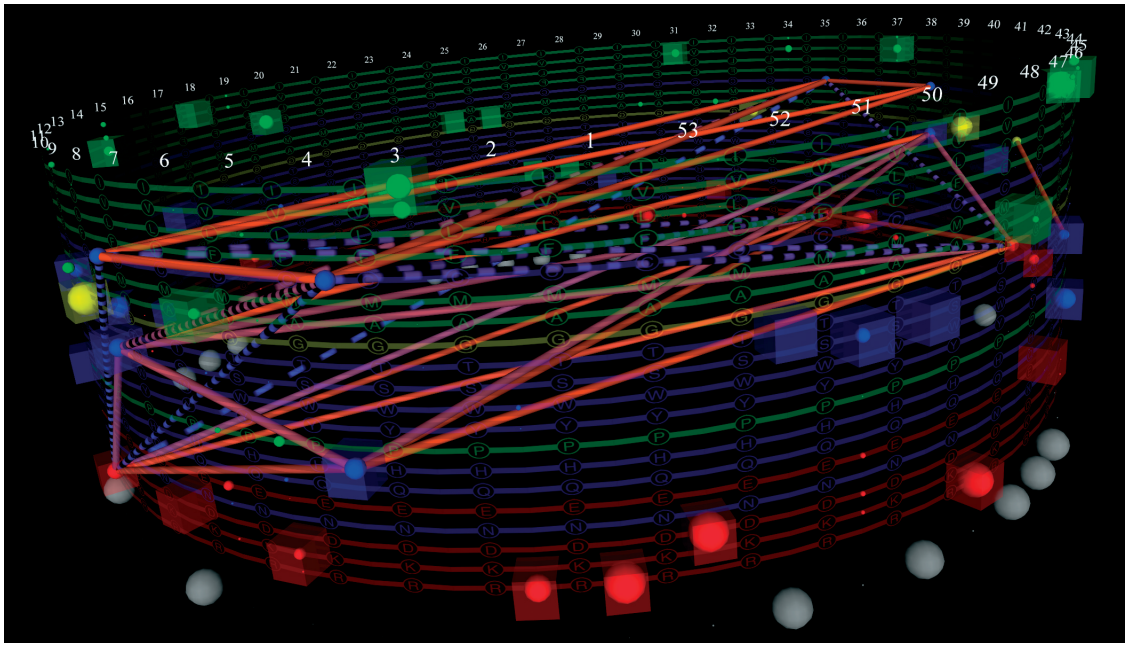

Figure 2. A StickWRLD diagram showing related positions in the ADK_Lid domain. The positively related cysteines at 4, 8, 35 and 38 stabilize the domain structure using bound zinc in a conformation analogous to a zinc finger, while the positively related histidine, serine, aspartic acid and threonine in the same positions stabilize the same domain structure using a network of hydrogen bonds. The amino acid identities are arranged vertically by their Kyte–Doolittle hydropathy score, and are colored using Branden and Tooze's classical coloring and grouping of residues by hydrophobic character. Consensus identities in each position are highlighted by a transparent unit cube. In a live VRML browser, this diagram is completely navigable and the viewer can rotate, move and zoom the 3D diagram to examine details of any portion.

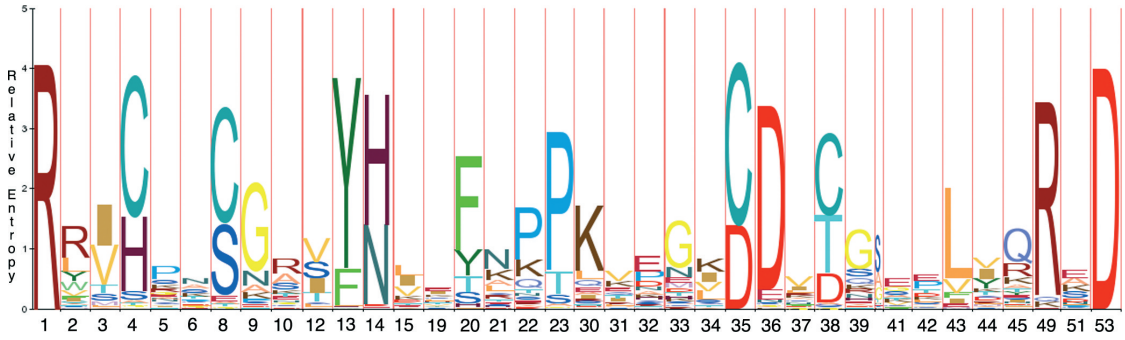

Figure 3. A LogoMat-M HMM-Logo visualizing the HMM defined by the Pfam ADK_lid training sequence set, renumbered to match the complete 299-member predicted family. The logo provides a convenient visualization of the identity probabilities at each position, and to what extent each position contributes to the information content of the model, but it does not suggest the apparent requirement for either of the alternate C4, C8, C35, C38 or H4, H8, R10, D35, T38, E41 motifs that are found in the actual sequences.

webserver itself) are generated from the entire family, and disguise the specific sequence requirements belonging to the subfamilies. The StickWRLD diagram indicates a strong positive relationship between the cysteines at 4 and 8 (and the coordinated pair at 35 and 38), as well as a strong positive relationship between histidine and serine at 4 and 8, and a strong negative relationship between these and the quadruple cysteines. Pfam provides HMMer-generated HMMs [http://hmmer.wustl.edu/] for the training sequence set, and assists the user in generating an HMM Logo (17) for visualizing the HMM properties. The HMM logo generated for ADK_lid, shown in Figure 3; however, does not suggest the interpositional requirements found in the sequences, and the HMM it represents would erroneously score a sequence that had H and S in the 4 and 8 positions, coupled with a pair of cysteines at 35 and 38, as a better match for the family than one containing aspartic acid and threonine (D and T) at 35 and 38. In fact, a HMM bootstrapped from Pfam's full 299-member ADK_lid family, will score a (possibly disallowed) C, T pair at 35, 38 better than either of the biologically relevant C, C or D, T pairs.

Additional study of the StickWRLD diagram suggests that the subfamily distinguishing features might not be simply the (H4, S8, D35, T38) or (C4, C8, C35, C38) quadruple, but rather that there are three residues on each loop that are coordinated. H4 and S8 also show a strong positive relationship to R at position 10, while C4, C8 is negatively related to R10, and prefers any other residue in that position. The HMM disguises this fact and actually down-scores C4, C8 motifs if they don't contain R10. Likewise on the opposite loop, D35, T38 shows a strong positive relationship to E at 41, while C35, C38 prefers other residues in that position. While Berry and Phillips (16) have shown that the cysteines in the 4-Cys variant of the domain are the only residues within 5 Å of the coordinated zinc, it seems likely that R10 and E41 have a structural role in the H4, S8, D35, T38 variant. Figure 4 shows the three-dimensional backbone structure of the Bovine ADK_lid domain (1AK2, an H, S, D, T variant), with the side chains for the H4, S8, R10 and D35, T38, E41 triplets shown. Clearly, R10 and E41 are appropriately oriented to participate in structural interactions with H4, S8, D35 and T38.

By reducing the stringency of the MAVL parameters (Figure 2 StickWRLD diagram was generated to show only residuals containing 20% or more of the total population, a

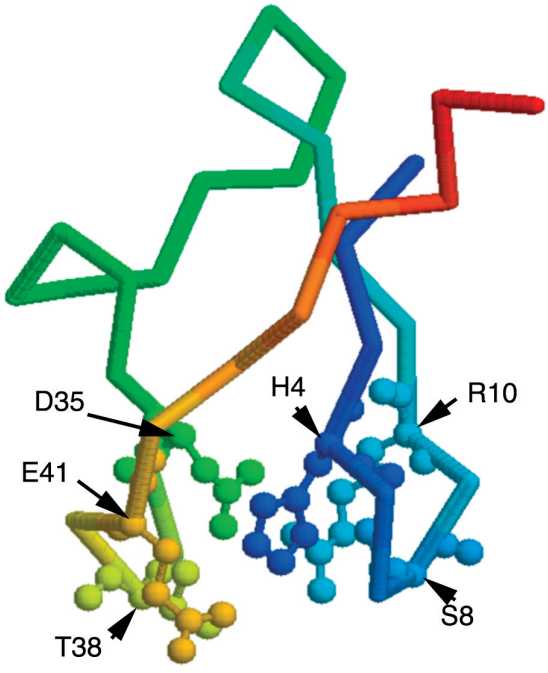

Figure 4. A RasMol rendition of the bovine ADK_lid structure. Residues 4, 8, 35 and 38 are mutated to cysteines in Gram-positive bacterial ADK_lid domains. Residues 10 and 41 are implicated in a structural capacity in variants lacking the cysteines, due to strong positive relationships between their identities, and those of the 4, 8, 35, 38 quadruple.

rather strict setting), it becomes apparent that there is a larger network of positive and negative effects that all display considerable statistical significance. All of these are situations where a Consensus, Weight-Matrix, or HMM model of a sequence family, at best discards statistically significant information about the characteristics of the sequence family, and more likely, produces results that overscore sequences with characteristics that are denied by the real family, and underscore sequences that should be accepted as family members.

CONCLUSION

Like MAVL/StickWRLD for NA, MAVL/StickWRLD for protein provides a rapid and convenient method for visually

surveying the limitations of Consensus, Weight-Matrix or HMM models that can be generated to describe a family of sequences. It highlights situations (such as that shown by ADK_lid) where a family may be better described by a pair of variant models rather than a single model that disguises important sequence requirements and limitations.

Although MAVL/StickWRLD was designed for examining statistically significant interpositional dependencies rather than compensatory substitutions that might be the cause of such dependencies, with use it has become increasingly clear that many of the significant links that are discovered, have probable structural relationships. This is interesting, because MAVL makes no attempt to filter near phylogenetic neighbors, or to specifically classify positions that display significant changes in properties between family members. To assist researchers who are interested in applying MAVL/StickWRLD to predicting protein folding constraints, StickWRLD now colors links based on the similarity of the related residues, based on the viewer's choice of amino acid size, hydropathy, composition or polarity. We are also exploring extensions that will group residues based on bulk properties and highlight regions with apparently compensatory changes, to simplify visual surveys for possible spatial relationships.

ACKNOWLEDGEMENTS

The Open Access publication charges for this article were paid by W.C.R.

Conflict of interest statement. None declared.

REFERENCES

1. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
2. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
3. Gregoret,L.M. and Sauer,R.T. (1993) Additivity of mutant effects assessed by binomial mutagenesis. *Proc. Natl Acad. Sci. USA*, **90**, 4246–4250.
4. Taylor,W.R. and Hatrick,K. (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.*, **7**, 341–348.
5. Afonnikov,D.A. and Kolchanov,N.A. (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res.*, **32**, W64–W68.
6. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
7. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
8. Ray,W.C. (2004) MAVL and StickWRLD: visually exploring relationships in nucleic acid sequence alignments. *Nucleic Acids Res.*, **32**, W59–W63.
9. Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
10. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
11. Kawashima,S. and Kanehisa,M. (2000) Aaindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.
12. Branden,C. and Tooze,J. (1999) *Introduction to Protein Structure*. Garland Publishing, Inc., New York, NY.
13. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
14. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
15. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
16. Berry,M. and Phillips,G.N.Jr (1998) Crystal Structures of *Bacillus stearothermophilus* Adenylate kinase with bound Ap_5A, Mg^{2+} Ap_5A, and Mn^{2+} Ap_5A reveal an intermediate lid position and six coordinate octahedral geometry for bound Mg^{2+} and Mn^{2+}. *Prot. Str. Func. Gen.*, **32**, 276–288.
17. Schuster-Böckler,B., Schultz,J. and Rahmann,S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.