Available online at www.sciencedirect.com
ScienceDirect
Insect Biochemistry and Molecular Biology 37 (2007) 754–760
Insect Biochemistry and Molecular Biology
www.elsevier.com/locate/ibmb

Review

Drosophila cuticular proteins with the R&R Consensus: Annotation and classification with a new tool for discriminating RR-1 and RR-2 sequences

Maria V. Karouzou[a], Yannis Spyropoulos[a], Vassiliki A. Iconomidou[a], R.S. Cornman[b], Stavros J. Hamodrakas[a], Judith H. Willis[b,*]

[a]*Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 157 01, Greece*
[b]*Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA*

Abstract

The majority of cuticular protein sequences identified to date from a diversity of arthropods have a conserved region known as the Rebers and Riddiford Consensus (R&R Consensus). This consensus region has been used to query the whole genome sequence of *Drosophila melanogaster*. One hundred one putative cuticular proteins have been annotated. Of these, 29 had been annotated previously, and for several their authenticity as cuticular proteins had been verified by protein sequence data from isolated cuticles or by localization of their transcripts in epidermis when cuticle synthesis was occurring. The original names have been retained, and the 72 newly annotated proteins have been given names beginning with Cpr followed by the chromosomal band in which the gene is located.

Proteins with the R&R Consensus can be split into three groups RR-1, RR-2 and RR-3, with some correlation to the type or region of the cuticle in which they occur. Previous classification was manual and subjective. We now have developed a tool using profile hidden Markov models that allows more objective classification. We describe the development and verification of the validity of this tool that is available at the cuticleDB website ⟨http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp⟩.

Keywords: Cuticle; Profile hidden Markov modeling

1. Introduction

The cuticle of arthropods is a composite of chitin and cuticular proteins. While chitin is a uniform polymer of *N*-acetylglucosamine, the protein component is made up of a multiplicity of cuticular proteins. A landmark discovery was made when Rebers and Riddiford (1988) recognized a common motif in some of the few cuticular protein sequences then available. What is remarkable is that only six sequences, of which only five were complete, formed the basis of what is known as the Rebers and Riddiford Consensus (R&R Consensus) that is present in 72% of the 519 cuticular protein sequences available in December, 2006 at cuticleDB (Magkrioti et al., 2004) ⟨http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp⟩. While that fraction may be biased because of the ease of using the consensus to define proteins, even prior to using any data from whole genome sequencing projects, 70% of the 139 cuticle protein sequences available in 2004 had the consensus (Willis et al., 2005). The consensus has now been found in cuticular proteins from 24 species of insects from seven orders and also from four crustaceans and two Chelicerata.

The original consensus was **G-*x*(8)-G-*x*(6)-Y**-*x*-A-*x*-E-*x*-**GY**-*x*(7)-P-*x*-P. The *x* is used to designate that a diversity of amino acids can be used in those positions. This original consensus has undergone some modifications, but the residues bolded and their spacing remain the hallmark of

*Corresponding author. Tel.: +1 706 542 0802; fax: +1 706 542 4271.
E-mail address: jhwillis@cb.uga.edu (J.H. Willis).

0965-1748/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2007.03.007

this region. The final Y can be replaced by F, and up to six amino acids, instead of the original five, can intervene between the first Y and the GY. Since its first mention, the consensus has been widely studied. First it was recognized that the consensus region could be extended N-terminal and frequently began with an aromatic triad (Y/F)-*x*-(Y/F)-*x*-(Y/F). This "extended R&R Consensus" is recognized as pfam00379 (IPR000618) and examples are in Fig. 1. Then it was appreciated that there were three distinct forms of the extended consensus that Andersen (1998, 2000) ultimately named RR-1, RR-2 and RR-3. Proteins of the RR-1 form were first isolated from soft (flexible) cuticles, while RR-2 proteins were associated with rigid (hard) cuticle. Later Andersen (2000) suggested that the RR-2 proteins would be secreted first and deposited in the pre-ecdysial (exo-) cuticle while RR-1 proteins might be secreted into the post-ecdysial (endo-) cuticle, although he lamented that the data were too sparse to be certain. A recent proteomics analysis was carried out on cuticle left behind after *Anopheles gambiae* had molted. Since there had been considerable digestion by molting fluid, one assumes that what remained would be exocuticle. Yet peptides from both RR-1 and RR-2 proteins were recovered (He et al., 2007). The RR-3 consensus has only been recognized in five sequences from postecdysial cuticle of insects plus sequences from other arthropod classes (Andersen, 2000). Obviously, more work is needed to learn the precise roles of these highly conserved classes of RR proteins and that will be aided by a simple and consistent way to distinguish among the classes.

The important lesson is the conservation of the R&R Consensus across the arthropods. Rebers and Riddiford (1988) suggested that the consensus "plays an important functional role in cuticular structure." Others elaborated by suggesting that it bound to chitin. Chitin binding has recently been probed by homology modeling (Hamodrakas et al., 2002; Iconomidou et al., 2005) and confirmed experimentally (Rebers and Willis, 2001; Togawa et al., 2004). β-pleated sheet is most probably the underlying molecular conformation of a large part of the extended R&R Consensus, especially the part which contains the R&R Consensus itself, and this conformation is likely involved in β-sheet/chitin–chain interactions of the cuticular proteins with the chitin filaments (Iconomidou et al., 1999, 2001, 2005; Hamodrakas et al., 2002).

Riddiford was involved in another major contribution to insect cuticle biology. Charles et al. (1997, 1998) identified a region at band 65A on the *Drosophila melanogaster* chromosome 3L that had genes for 12 cuticular proteins of the RR-1 type and a pseudogene. Two pairs of genes were very similar, and their copy number varied among strains. Their sophisticated analysis of the region identified features that might have contributed to the formation of this cluster. In addition, they verified the N-terminal regions of several of the predicted proteins by sequencing proteins eluted from gels and used Northern analyses to learn when each gene was expressed. Clustered cuticular protein genes were first discovered by Snyder et al. (1982) also in *D. melanogaster*, and it was two of their four "larval" cuticular proteins along with two *Manduca* proteins and another *D. melanogaster* protein, Pcp (discussed later), that contributed to the recognition of the R&R Consensus.

It seemed appropriate to honor Professor Lynn Riddiford on the occasion of her 70th Birthday, by using the R&R

Fig. 1. (A) Multiple alignment of the RR-1 consensus from 14 RR-1 proteins of *D. melanogaster* used to train the RR1 HMM. (B) Multiple alignment of the RR-2 consensus from 9 RR-2 proteins of *D. melanogaster* used to train the RR2 HMM. The original R&R Consensus is highlighted for the first sequence in each panel. Additional information about these proteins can be found in Table 1 (see Appendix A).

Consensus to identify all the cuticular proteins in the fully sequenced genome of *D. melanogaster*. This paper reports on that analysis as well as the development and use of a tool based on profile hidden Markov models (HMMs) capable of discriminating between RR-1 and -2 cuticular proteins so they can be readily classified.

2. Methods

2.1. Annotation

The whole genome sequence of *D. melanogaster* has been subjected to several rounds of annotation and is now in version 5.1. We identified genes that might code for cuticular proteins with the R&R Consensus by searching Ensembl 〈http://www.ensembl.org/Drosophila_melanogaster/index.html〉 and FlyBase 〈http://flybase.bio.indiana.edu/〉 with pfam00379 = IPR00618—a HMM that recognizes all classes of proteins with the R&R Consensus. One hundred three sequences identified were then checked to verify that they had an R&R Consensus and a signal peptide. There is a vast array of ESTs and cDNAs available for *D. melanogaster* and each predicted gene was also examined to verify that it was consistent with these available transcripts. Two sequences that came up during the initial survey (CG15756 and CG15515) did not appear to code for cuticular proteins; two others (CG13670 and CG31878) could not be properly annotated. Six predicted sequences required further annotation (see Section 3).

The region corresponding to chromosomal band 65A contains many genes for cuticular proteins and the number varies among strains (see Section 1). Hence the relevant region of AE014296.4 (6,115,000–6,155,000) was translated in all 6 reading frames and all putative cuticular protein sequences were identified with manual searching for signature regions. Two previously unrecognized genes for proteins with the R&R Consensus were recognized, Cpr65Ax1 and Cpr65Ay; there were two cDNAs that corresponded to the latter.

Assignment of RR class was made manually and by using the HMM described in this paper.

2.2. Naming protocol

Twenty-nine of the *D. melanogaster* RR proteins had already been named. Five used the designation LCP, one was named for its appearance in the pupa (Pcp), while three had names related to function [resilin, cry, l(3)mbn]. Names for the rest came from the chromosomal band in which each gene resided. Four genes were found to be induced by a pulse of 20-hydroxyecdysone and were named ecdysone-dependent gene (Edg) followed by the chromosomal band; two coded for proteins with the R&R Consensus (Fechtel et al., 1988; Apple and Fristrom, 1991). While Edg84A was correctly mapped, ''Edg78E'' is now reported by FlyBase to reside in band 78C. There are 7 RR genes associated with Band 84A, named Ccp84A (a–g) (Kaufman et al., 1990). The next study to identify multiple genes within one band, combined two previous conventions, calling them Acp65Aa for the one expressed in adults, and Lcp65A (a–g) for the others (Charles et al., 1997). Two pairs of very similar genes received the designation b1/b2 and g1/g2. That same study also identified a g3 in another strain.

We have adopted a simple nomenclature building on precedents described above and also incorporating recognition that the proteins have the R&R Consensus. All the newly identified RR genes were named Cpr followed by the band in which they occur. If multiple genes are present in a single band, the band name is followed by a,b, etc. To avoid confusion in the 65A region, we named the new bands, Cpr65A (u–z). The designation Cpr was chosen because that nomenclature is already in use for the proteins in *A. gambiae* that have the R&R Consensus (He et al., 2007), although there three capital letters (CPR) are used. Current rules for *Drosophila* nomenclature allow one capital letter.

Table 1 (see Appendix A) lists all 101 *D. melanogaster* genes that code for proteins with the R&R Consensus using the new nomenclature along with their RR class and synonyms.

2.3. Design of profile hidden Markov models

In order to facilitate the identification of RR-1 and RR-2 consensus regions, we used the package HMMER 2.3.2 (Eddy, 1998) to construct profile HMMs. Such models are statistical models of multiple sequence alignments that capture position-specific information and complement standard pairwise comparison methods for large-scale sequence analysis. For each consensus column of the multiple alignments, a 'match' state models the distribution of residues allowed in the column. An 'insert' state and 'delete' state at each column allow for insertion of one or more residues between that column and the next, or for deleting the consensus residue. Profile HMMs are strongly linear, left–right models. The probability parameters in a profile HMM are usually converted to additive log–odds scores before aligning and scoring a query sequence (Barrett et al., 1997). The scores for aligning a residue to a profile match state are, therefore, comparable to the derivation of BLAST or FASTA scores.

2.4. Datasets used for training and evaluating the methods

We chose sequences from only one species, *D. melanogaster*, in order to eliminate the probability of using homologous proteins from related species. This would constitute a factor of bias for the Markov models. The profile HMM that corresponds to type RR-1 proteins was built using the multiple alignment (Fig. 1A) of the RR-1 regions of 14 RR-1 proteins. The training dataset with 9 RR-2 proteins is aligned in Fig. 1B.

The proteins used for the initial test datasets and the final tests were taken from Figs. 1 and 2 of Willis et al. (2005) where the extended consensus regions of proteins that fit patterns for RR-1 and RR-2 proteins had been aligned. Protein sequences and additional annotation are available at cuticleDB 〈http://bioinformatics.biol.uoa.gr/cuticleDB〉.

They can also be retrieved via Entrez 〈http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi〉 using the Uniprot AC number that is provided in the Tables. In some cases, several AC numbers are available for a single protein, but we have simplified the Tables by using only one.

Two initial test datasets, one for each HMM, were used, so that we could calculate the cutoffs. The test dataset for RR-1 HMM consists of 35 RR-1 proteins that are not included in the training set, 44 RR-2 and 5 RR-3 proteins (Table 2 (see Appendix A)). The dataset used to test RR-2 HMM is composed of 49 RR-1 (35 included in the test dataset for RR-1 HMM plus the 14 proteins that were used to train the RR-1 HMM), 35 RR-2 proteins that are not included in the training set and 5 RR-3 proteins (Table 3 (see Appendix A)). In Tables 2 and 3 (see Appendix A) we provide the type based on prior manual annotation (RR-1, RR-2 or RR-3) and the score and *e*-value that were produced when running each protein against the respective HMM.

In Table 1, the hits of each HMM which can be used in order to validate the discriminative capability of each HMM are summarized.

In order to evaluate whether the models are fitted or not, standard measures of the predictive performance of the models were calculated from the test datasets for a range of 10 units of score: Precision = TP/(TP + FP), Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP) and Accuracy = (Sensitivity + Specificity)/2, where TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively (Tables 2 and 3). The cutoff was estimated by plotting sensitivity and specificity against the different cutoffs and finding the point where sensitivity and specificity meet (Figs. 2A and 2B). For RR-1 the cutoff is 35.0 and for RR-2 37.5. In addition, we checked the scores of true positives and false positives of each profile HMM and found out that they are not overlapping (Table 4).

Table 1
Validation of the discriminative capability of the HMMs

	Test set		
	RR-1	RR-2	RR-3
RR-1 HMM	37/37	0/44	0/5
RR-2 HMM	0/51	35/35	0/5

Table 2
Calculation of TP, TN, sensitivity and specificity of the RR-1 HMM based on the results obtained for the test dataset for every 10 units of score

Cutoffs	TP	TN	Sensitivity	Specificity
10	35	45	1.0000	0.9184
15	35	48	1.0000	0.9795
20	35	48	1.0000	0.9796
25	35	49	1.0000	1.0000
30	35	49	1.0000	1.0000
35	35	49	1.0000	1.0000
40	35	49	1.0000	1.0000
45	35	49	1.0000	1.0000
50	32	49	0.9143	1.0000

Table 3
Calculation of TP, TN, sensitivity and specificity of the RR-2 HMM based on the results obtained for the test dataset for every 10 units of score

Cutoffs	TP	TN	Sensitivity	Specificity
10	35	44	1.0000	0.7857
15	35	49	1.0000	0.8750
20	35	51	1.0000	0.9107
25	35	56	1.0000	1.0000
30	35	56	1.0000	1.0000
35	35	56	1.0000	1.0000
40	35	56	1.0000	1.0000
45	35	56	1.0000	1.0000
50	35	56	1.0000	1.0000
55	34	56	0.9714	1.0000

3. Results and discussion

3.1. Annotation and description of the D. melanogaster proteins with the R&R Consensus

Most of the genes recognized by IPR000618 appear to be correctly annotated. The six exceptions are: Cpr49Aa (CG30045), Cpr64Ab (CG15007), Cpr66Ca (CG7072), Cpr73D (CG9665), Cpr76Bc (CG9295-PB), Cpr76Bd (CG9299-PA). Two RR genes, Cpr65Ax1 and Cpr65Ay, had not been annotated. Details of each revision and new annotation are provided in a footnote to Table 1 (see Appendix A).

We have identified 101 genes for cuticular proteins with the R&R Consensus. Of these, 55 are RR-1, 37 are RR-2 and 9 could not be classified using the HMM tool. Three of these plus two other with low HMM predictions of RR-2 appear by manual annotation and clustering to be RR-3. The median size of the processed (mature) proteins is 145 amino acids (~15,000 D) with only 15% being longer than 300. The range is from 76–1211. One of these large proteins, Cpr73D is of particular interest because it appeared to be an ortholog of AgamCPR144, a protein with 3 RR regions. These regions are marked on the sequence for Cpr73D in the footnote to Table 1 (see Appendix A).

Our annotation further supports the finding of intraspecific variation in CPR gene number in *D. melanogaster* previously reported by Charles et al. (1997). That study identified 12 genes plus a pseudogene within the 65A region, whereas our annotation of the complete genome sequence, of a different strain, includes 19 genes with the R&R Consensus in the same region. It is interesting that the genes at the outer borders of this sequence [l(3)mbn-RB and Cpr65Az] are longer than the others.

This study also confirms two important precedents with respect to CPR gene architecture. There was, until now, no

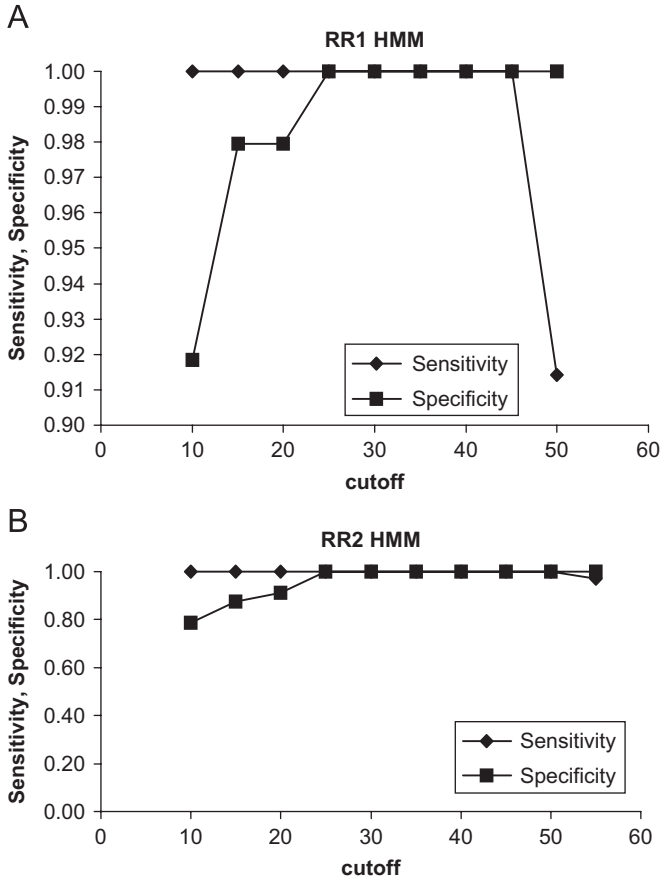

Fig. 2. (A) Plot of sensitivity and specificity against the different cutoffs, in order to find the cutoff for RR-1 HMM. (B) Plot of sensitivity and specificity against the different cutoffs, in order to find the cutoff for RR-2 HMM.

Table 4
Scores for true positives and false positives

	RR-1 HMM	RR-2 HMM
Lowest score for proteins of the same type	45.9	51.0
Highest score for proteins of different type	21.2	22.8

direct evidence for a gene with the R&R Consensus having more than a single transcript. CG8502, however, was predicted to have two transcripts (Cpr49Ac-RA and Cpr49Ac-RC) and we have identified EST support for each. Many years ago it was recognized that Pcp was found in an intron of a gene for a constitutive enzyme, oriented in the opposite direction (Henikoff et al., 1986). Our analysis yielded two other genes (Cpr51A and Cpr56F), each associated with another gene. In each case the unrelated gene lies within the first intron of the cuticular protein gene and in the opposite orientation. All four of these genes are supported by numerous ESTs, indicating that all four are highly expressed. A hint that they might have different temporal and spatial patterns of expression is that the ESTs came from different libraries.

3.2. Design of HMM tool to identify and classify cuticular proteins with the R&R Consensus

We created a dataset that consisted of all the RR-1 and RR-2 proteins that were deposited in cuticleDB 〈http://bioinformatics.biol.uoa.gr/cuticleDB〉 as of December, 2006, excluding those that were used in the training sets. This dataset included 132 RR-1 proteins and 173 RR-2 proteins.

The RR-1 HMM recognized 127 proteins previously characterized by manual analysis as RR-1 with a score >35 (96.21%) and none of the RR-2 proteins (Table 4 (see Appendix A)). The RR-2 HMM correctly predicts 166 RR-2 proteins with a score >37.5 (95.95%) and none of the RR-1 proteins (Table 5 (see Appendix A)). Hence, both HMMs reliably recognize proteins of the same RR class by assigning high scores, while proteins of a different type receive low scores.

The two profile HMMs can be used as predictive tools for proteins that bear the R&R Consensus, without knowing its specific type (RR-1 or RR-2). Our method is freely available at 〈http://bioinformatics.biol.uoa.gr/cuticleDB/hmmfind_form.jsp〉. The user may submit a sequence or a collection of sequences in FASTA format and has the option of choosing to run the profile HMM for RR-1, RR-2 or both. It is also possible to select the score cutoff and the *e*-value cutoff. By default, we have set the score cutoff at 0.0 and the *e*-value cutoff at 5.0*e*-07. The application returns the protein name, its type (RR-1 or RR-2), the score and the *e*-value. There is also a link to the hmmpfam output, where the user can view the alignments of top-scoring domains. This link is essential to properly analyze those rare proteins with more than a single consensus region.

To test this new feature on cuticleDB, we subjected all of the annotated RR proteins from *D. melanogaster* to the analysis as a single FASTA file. The values obtained are in Table 1 (see Appendix A). The consensus region was recognized and characterized even for Cpr76Bd where it comprises less than 6% of the total sequence. Only 9 proteins could not be assigned, and three of these are RR-3 (see below).

3.3. Relationships among D. melanogaster genes with the R&R Consensus using neighbor-joining trees

Comparisons were made at the amino acid level among the cuticular proteins we had identified using neighbor-joining trees drawn using Mega3 〈www.megasoftware.net〉 (Kumar et al., 2004). For this purpose, only the extended consensus region was used because it can be aligned across CPR genes. For RR-2 proteins, this was a standard 63 amino acids from two amino acids before the aromatic triad to 8 amino acids after the final G(F/Y). An extended consensus is not as readily defined for RR-1 proteins because there is less sequence conservation and indels are common near the center of the consensus. The ends of the region can be aligned between groups, however.

All but one of the genes identified as RR-2 by the HMM tool formed a distinct clade of the neighbor-joining tree, which had a bootstrap support of 37%. Subsets of proteins from physically linked genes were clustered at lower bootstrap values, but these clusters appear to be biologically informative. In order to illustrate such clusters, we present a condensed tree with nodes supported by at least 20% (Fig. 3). Of the nine proteins that were not classified by the HMM tool, three (Cpr67B, Cpr72Ea, Cpr72Eb) appeared by manual annotation to be RR-3 proteins and were clustered in the tree with two proteins assigned as RR-1 (Cpr72Ec, Cpr92F, but which also matched criteria for RR-3. The node joining these 5 proteins had 42% support. Four of the remaining 6 unclassified genes fell between the RR-1 and RR-2 classes [l(3)mbn, Cpr51A, Cpr65Ay, Cpr97Eb]. The remaining two were embedded among the RR-1 proteins (Cpr12A and Cpr60D). Thus, genes that were not classified by the HMM tool appear to have different evolutionary origins; some are intermediate between RR-1 and RR-2 whereas others appear to be independently derived deviations from the RR-1 consensus. This phylogenetic perspective further demonstrates the utility of the HMM tool. The tree also shows that the large number of cuticular protein genes coding for proteins with the R&R Consensus seems to be due to duplications that occurred long ago. This conclusion is suggested by the absence of large clusters of highly similar genes, as only four pairs of genes had greater than 90% identity. In contrast, *A. gambiae* has large clusters of closely related genes (He et al., 2007; Cornman, Dunn and Willis unpublished observations).

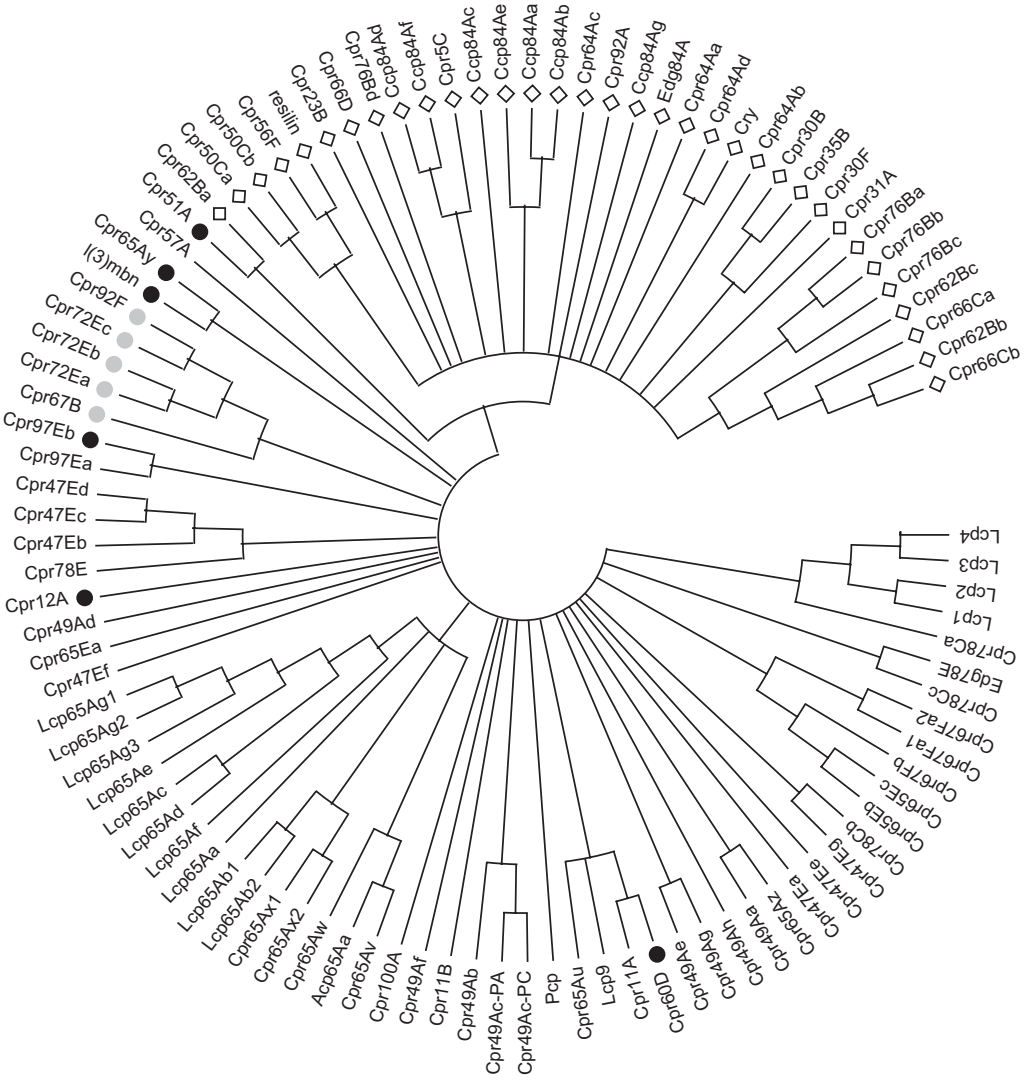

Fig. 3. Neighbor-joining tree of annotated *Drosophila melanogaster* CPR proteins (Cpr73D omitted). The tree is condensed to show only branches with 20% bootstrap support or higher. This value was chosen because clades supported at this level are also supported by chromosomal position, i.e. they are physically clustered and there is independent evidence of gene duplication and gene conversion in such regions (Charles et al., 1997). The tree was generated with MEGA3 (Kumar et al., 2004) using the JTT cost matrix and pairwise deletion of indels. Bootstrap support is based on 1000 resampled data sets. Symbols before gene names indicate RR type: open diamonds = RR-2; gray circles = RR-3; solid circles = not assigned by HMM model; all others are RR-1.

3.4. Summary

This paper presents names and correct annotation for the cuticular proteins of *D. melanogaster* that bear the R&R Consensus. It also introduces a tool that will allow ready classification of most proteins with the consensus into RR-1 or RR-2. We expect that predictions from our method will be useful for researchers, as well as for bioinformatics analyses of published proteomes. Ultimately, the accumulated information coupled with more extensive data on temporal and spatial expression of the proteins with the R&R Consensus should inform us of the significance of differences in the consensus and in the flanking regions.

Acknowledgments

We thank Rachel Drysdale (Cambridge, UK) and Margo Roark (Cambridge, MA) from FlyBase for guidance on naming and annotating *Drosophila* genes and Aaron Emmons for assistance. We acknowledge the support from the National Institutes of Health (AI55624) to JHW for the work conducted at the University of Georgia.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ibmb.2007.03.007.

References

Andersen, S.O., 1998. Amino acid sequence studies on endocuticular proteins from the desert locust, *Schistocerca gregaria*. Insect Biochem. Mol. Biol. 28, 421–434.

Andersen, S.O., 2000. Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. Insect Biochem. Mol. Biol. 30, 569–577.

Apple, R.T., Fristrom, J.W., 1991. 20-Hydroxyecdysone is required for, and negatively regulates, transcription of *Drosophila* pupal cuticle protein genes. Dev. Biol. 146, 569–582.

Barrett, C., Hughey, R., Karplus, K., 1997. Scoring hidden Markov models. CABIOS 13 (2), 191–199.

Charles, J.-P., Chihara, C., Nejad, S., Riddiford, L.M., 1997. A cluster of cuticle protein genes of *Drosophila melanogaster* at 65A: sequence, structure and evolution. Genetics 147, 1213–1226.

Charles, J.-P., Chihara, C., Nejad, S., Riddiford, L.M., 1998. Identification of proteins and developmental expression of RNAs encoded by the 65A cuticle protein gene cluster in *Drosophila melanogaster*. Insect Biochem. Mol. Biol. 28, 131–138.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Fechtel, K., Natzle, J.E., Brown, E.E., Fristrom, J.W., 1988. Prepupal differentiation of *Drosophila* imaginal discs: identification of four genes whose transcripts accumulate in response to a pulse of 20-hydroxyecdysone. Genetics 120, 465–474.

Hamodrakas, S.J., Willis, J.H., Iconomidou, V.A., 2002. A structural model of the chitin-binding domain of cuticle proteins. Insect Biochem. Mol. Biol. 32, 1577–1583.

He, N., Botelho, J.M.C., McNall, R.J., Belozerov, V., Dunn, W.A., Mize, T., Orlando, R., Willis., J.H., 2007. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. Insect Biochem. Mol. Biol. 37, 135–146.

Henikoff, S., Keene, M.A., Fechtel, K., Fristrom, J.W., 1986. Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell 44, 33–42.

Iconomidou, V.A., Willis, J.H., Hamodrakas, S.J., 1999. Is β-pleated sheet the molecular conformation which dictates formation of helicoidal cuticle? Insect Biochem. Mol. Biol. 29, 285–292.

Iconomidou, V.A., Chryssikos, G.D., Gionis, V., Willis, J.H., Hamodrakas, S.J., 2001. "Soft"-cuticle protein secondary structure as revealed by FT-Raman, ATR FT-IR and CD spectroscopy. Insect Biochem. Mol. Biol. 31, 877–885.

Iconomidou, V.A., Willis, J.H., Hamodrakas, S.J., 2005. Unique features of the structural model of 'hard' cuticle proteins: implications for chitin–protein interactions and cross-linking in cuticle. Insect Biochem. Mol. Biol. 35, 553–560.

Kaufman, T.C., Seeger, M.A., Olsen, G., 1990. Molecular and genetic organization of the Antennapedia gene complex *of Drosophila melanogaster*. Adv. Genet. 27, 309–362.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for Molecular evolutionary genetics analysis and sequence alignment briefings. Bioinformatics 5, 150–163.

Magkrioti, C.K., Spyropoulos, I.C., Iconomidou, V.A., Willis, J.H., Hamodrakas, S.J., 2004. cuticleDB: a relational database of Arthropod cuticular proteins. BMC Bioinformatics 5, 138–143.

Rebers, J.F., Riddiford, L.M., 1988. Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J. Mol. Biol. 203, 411–423.

Rebers, J.E., Willis, J.H., 2001. A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem. Mol. Biol. 31, 1083–1093.

Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J., Davidson, N., 1982. Cuticle protein genes of *Drosophila*: structure, organization and evolution of four clustered genes. Cell 29, 1027–1040.

Togawa, T., Nakato, H., Izumi, S., 2004. Analysis for the chitin recognition mechanism of cuticle proteins from the soft cuticle of the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. 34, 1059–1067.

Willis, J.H., Iconomidou, V.A., Smith, R.F., Hamodrakas, S.J., 2005. Cuticular proteins. In: Gilbert, L.I., Iatrou, K., Gill, S.S. (Eds.), Comprehensive Molecular Insect Science, vol. 4. Elsevier, Oxford, UK, pp. 79–110.