Bioinformation
www.bioinformation.net

open access
Hypothesis

Classification of DNA sequences based on thermal melting profiles

Edward Reese[1], Vishwanathan V. Krishnan[1,2,3,*]

[1]Department of Chemistry, California State University Fresno, Fresno CA 93740; [2]Department of Pathology & Laboratory Medicine, University of California Davis Medical Center, Sacramento, CA 95837 ; [3]Department of Applied Science, University of California, Davis, CA, 95616. V. V. Krishnan -Email id : vvkrishnan@ucdavis.edu. Tel. +530 746-8220. Fax. +559 278-4402

Received December 03, 2009; accepted December 13, 2009; published April 30, 2010

Abstract:
A new classification scheme based on the melting profile of DNA sequences simulated thermal melting profiles presented. This method was applied in the classification of (a) several species of mammalian - β globin and (b) α-chain class II MHC genes. Comparison of the thermal melting profile with the molecular phylogenetic trees constructed using the sequences shows that the melting temperature based approach is able to reproduce most of the major features of the sequence based evolutionary tree. Melting profile method takes into account the inherent structure and dynamics of the DNA molecule, does not require sequence alignment prior to tree construction, and provides a means to verify the results experimentally. Therefore our results show that melting profile based classification of DNA sequences could be a useful tool for sequence analysis.

Key words:DNA, hybridization, melting profiles, classification.

Background:

The DNA double helix has more information built into its structure, that are both local such as variation in base pairing interactions and stacking interactions, as well as long range such as dynamic superhelical stress **[1-3]**. These interactions are responsible for the physical chemistry of the sequences and they are reflected in the thermodynamic properties such as the melting temperature. Therefore, sequences that have a high homology are expected to have similar thermodynamic parameters such as the melting temperature (T_m). Melting of a DNA molecule involves the denaturation the double-stranded DNA molecule into two single strands and it is the reverse process of hybridization. The denaturation process can be affected by many means such as an increase in temperature or denaturant concentrations **[4]**. The melting curve represents the denaturation process as a function of increasing sample temperature. Experimentally, the melting profile can be monitored by optical techniques such as absorption and fluorescence microscopy as the interactions among stacked bases cause a decrease in UV absorption. Melting of double-stranded DNA at elevated temperatures involves breaking the hydrogen bonds of the base pairs and a decrease of base stacking. This results in an increase in UV absorption, a hyperchromicity, which can be measured with a spectrophotometer **[4]**.

Melting curve analysis has been used in many applications such as the detection of single nucleotide polymorphisms (SNPs) **[5-7]** and has recently been proposed as an approach to DNA sequencing **[8]**. DNA melting profile analysis has also been used in many clinical research applications **[9]**; these include genotyping **[10-13]**, mutation scanning **[14, 15]** and simultaneous genotyping and mutation scanning **[16-18]**. Experiments based on melting profiles have also been used as a rapid, economical means of screening close relatives for transplant compatibility **[19]**. The melting behavior and melting temperature of oligonucleotides can be predicted by a wide range of thermodynamic models which assumes that the stability of a DNA duplex depends on the identity and orientation of the neighboring base pairs **[5-7, 20-25]**. The idea of using the thermal stability, in particular the melting temperature to differentiate between DNA sequences was originally suggested in the pioneering work by King and Wilson **[26]** and later followed by others **[26-29]**. King and Wilson used the nucleic acid hybridization melting temperature to quantify the resemblance between human and chimpanzee genes **[26]**. The difference in melting temperature between the reannealed human DNA and human-chimpanzee hybrid DNA is about 1.1° C, and that to the sibling species of *Drosophila*, congeneric species of mice and congeneric species of *Drisophila* are 3° C, 5° C and 19° C, respectively. Higher the difference in the T_m, larger the evolutionary distance between the species. As longer DNA sequences tend to have several localized melting events **[30, 31]**, the melting profile of the DNA sequence has additional information **[9]**. In one of the early works, Schmid and Marks **[28]** demonstrated the use of DNA hybridization as a guide to phylogeny using a model system of heteroduplexes formed between human β globin CDNA and four β -like globin genes isolated from a different species (chimpanzee).

In this study, we present a simple method for classifying the nucleotide sequences using simulated melting profiles. We demonstrate the utility of this method in β -globin and gene clusters of MHC class II α-chain proteins across multiple mammalian species. Comparison of the melting profile generated phylogenetics with that of the conventional sequence based approach reproduces many of the major features, but do not show a perfect match. The major advantage of this method is that it provides a way to verify phylogenetics constructed only from the DNA sequence and the molecular evolutionary process experimentally.

Methodology:

The first exon of the β -globin gene is often used as a standard example in many DNA-based graphing methods **[32]** is used as the first example. The gene family of β -globin varies between 86-105 bases and has a significant biological role in oxygen transport. β -globin gene across the 11 species is listed in **Table 1 (see supplementary material).** The class II MHC α-chain sequences were obtained from the original reference of Takahashi and co-workers **[33] (see supplementary material for data).** The average length of the class II MHC α-chain is 612 and a total of 31 different species were used for the analysis. The melting profile was separately calculated for each individual sequence using the MELTSIM program **[34, 35]**. The melting profiles were calculated from 60°C to 120°C in steps of 0.1°C using the default setting (75mM of NaCl). Classification of the calculated profiles were performed using the Euclidian distance measure, unweighted arithmetic average for clustering with a 10 fold bootstrapping for revalidation. Clustering and validation of the profiles were performed using a combination of codes written in Matlab **[36]** and R **[37, 38]**. For the same DNAs, sequence-based evolutionary trees were constructed using MEGA **[39]**. All phylogenetic output files were generated in newick tree format (http ://evolution.genetics.washington.edu/phylip/newicktree.html), and visualized using Treedyn **[40]**.

Results:

The gene family of β-globin across the 11 species are listed in **Table 1** along with the percentage of the various bases, GC content as well the estimated T_m, for each sequence. **Figure 1** shows the simulated melting profile using the program MELTSIM of the denaturation process (θ vs. temperature) (in **Figure.1a**) for all the 11 sequences and its first derivative dθ/dT (in **Figure. 1b**). The population value 0.5 and -0.5 represents the double-helical and denatured (single strand) DNAs, respectively and the melting temperature (T_m) is defined to be where these populations are equal.

The first derivative of the melting profile is more illustrative of the process of denaturation, as the melting temperature is represented as a peak (dθ/dT). Each derivative profile shows peak value that corresponds to the T_m and additional features such the width, shape and other low intensity peaks are manifestations of the sequence composition **[41, 42]**. As β-globin is represented by a relatively small number of bases such distinct features are not pronounced, except for the opossum (highlighted with dark curve in **Figure 1**). The DNA sequence of opossum has the lowest of the GC content of the sequences and consequently has the lowest T_m **[43]**. Melting profile based phylogenetics of β -globin is shown in **Figure 2a**, while the similar profile using the sequence is shown in **Figure 2b**. Most of the tree structures are reproduced in the melting profile based phylogenetics, with few notable differences. In both approaches the sequence of Opossum is clearly differentiated and the relative distances between Rat, Lemur and Gallus are also reproduced. Sequence based approach shows the cluster of Human, Chimpanzee and Gorilla, while the melting profile based approach keeps only the cluster of Chimpanzee and Gorilla together.

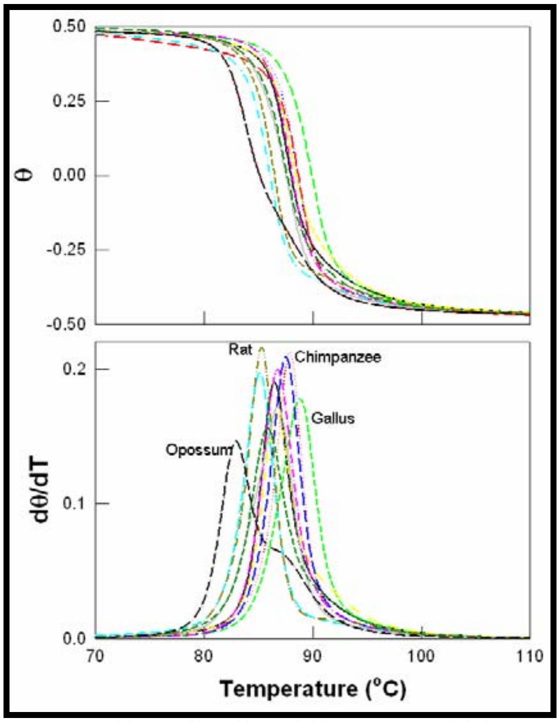

Figure 1: Melting profiles of beta-globin genes: (a) and their respective first derivatives (b). Some of the sequences are labeled in (b) and the melting profile of Opossum is shown in dark lines.

The shape of the melting curve and the melting temperature (T_m) are sensitive to the salt concentration in the sample **[44]**. All the calculations in figure 1 are performed with the salt concentration of 75mM. To highlight the effect of salt concentration on the melting profile, the β-globin sequence of opossum that shows additional features in the melting profile is simulated as a function of salt concentration. **Figure 3a** shows the plot of the profiles and the change in the T_m is shown in **Figure 3b**. Increase in salt concentration increases T_m (70°C at 0.01M to 98°C at 0.5M) and follows the empirical relationship between T_m vs. log(Concentration) **[44]**. Increasing the salt concentration also drastically changes the melting profile such as drastic loss of fine features that represent local melting events. Both melting profile and the T_m are important to differentiate the sequences in determining the phylogenetics and therefore optimal concentration of the salt is expected to be critical.

Figure 4 show the melting profiles and molecular phylogenetics, respectively for the genes of class II MHC α-chain genes **[33]**. The origins and divergence times of mammalian class II MHC genes have been studied in detail by Takahashi et al **[33]** and the data presented here is one of the subgroups of gene clusters studied elsewhere **([33])**. The sequence lengths of class II MHC α-chain genes are much larger than that of β -globin. Total of 31 different sequences and sequence length close to 612 for most sequences (see supporting material for the list genetic sequences) were used. As expected the melting profiles of these genes are complex suggesting the presence of additional features that could be useful to differentiate one sequence from the other. Original sequence based analysis of class II MHC α-chain genes showed four sequences (Zebra fish A1, A2 and A3 and shark) belong to an out-group and the melting temperature analysis **(Fig. 4)** also reproduces the same result. Mammalian class II MHC genes are clustered into four major groups, DRA, DPA, DQA and DNA **[33]**. Melting temperature based phylogeny is able to reproduce three of the four clusters only with overlaps between the closer clusters DQA and DPA.

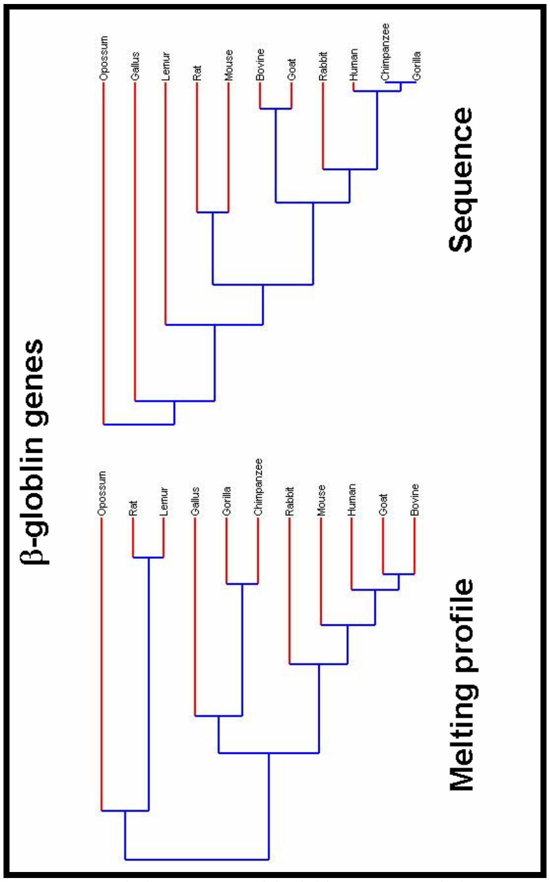

Figure 2: Evolutionary trees of β-globin genes: Evolutionary trees constructed using the thermal melting profiles (left) and that from the respective gene sequence after sequence alignment.

Discussion:

With the advent of new molecular biology techniques including sophisticated cloning, sequencing and monitoring genomes allows the characterization of the single species without need for cross-hybridization techniques. For example Liu *et al.* **[45]** have applied a melting map calculation to the complete human genome. (http://meltmap.uio.no). DNA melting curves analysis is a valuable technique in sequencing, differentiate between coding and non-coding regions **[46-48]**, genotyping **[46-48]** in the design of oligonucleotide probes in microarray experiments **[46-48]** and clinical applications **[19]** Classification of the profiles presented in this paper adds to an additional dimension to the power of melting profile analysis.

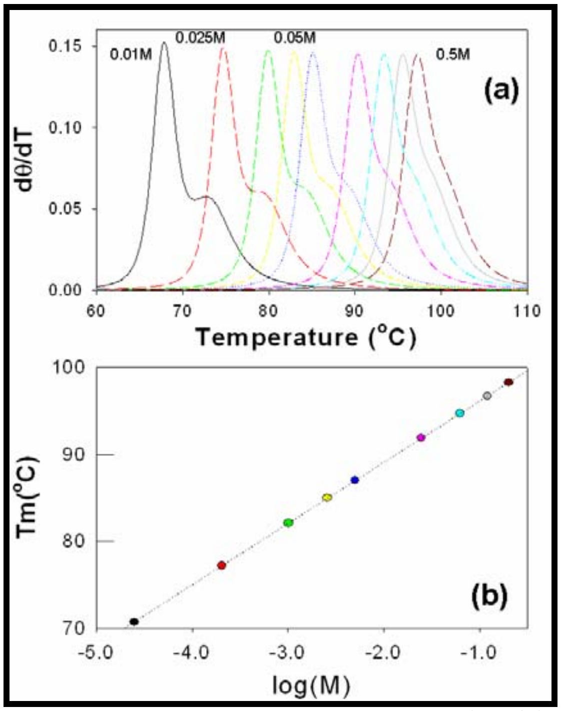

Figure 3: Temperature effect of Melting profiles: Plots of the melting profiles of the Opossum sequence of the β-globin gene family as a function of salt concentration (a). The plot of the melting temperature (T_m) as a function of the log (concentration) follows a liner relationship (b).

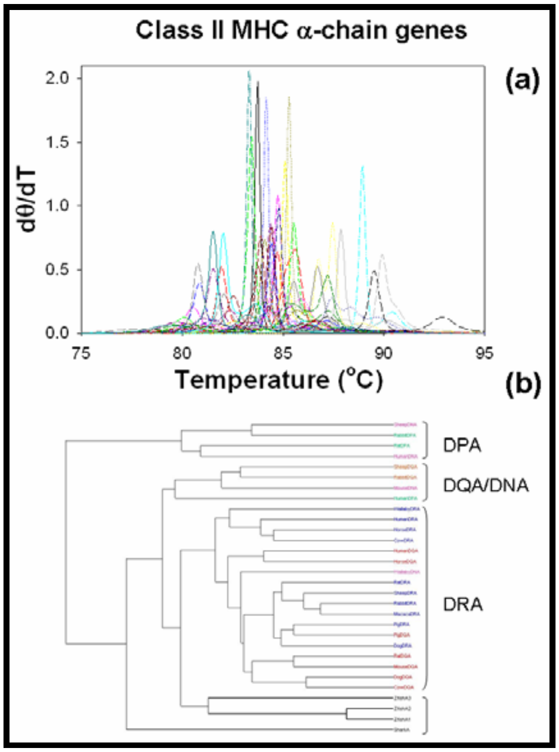

Figure 4: Application to Class II alpha-chains of MHC sequences: (a) Melting profiles of the 31 class II alpha-chains of MHC proteins simulated using MELTSIM. (b) Melting profile derived molecular evolutionary tree of class II alpha-chains of MHC proteins. DPA, DQA, DNA and DRA refer to the gene clusters originally **[33].**

Salt concentration is expected to play a significant role on the applicability of DNA melting based differentiation between the sequences **(Figure 2).** The hypochromicity of DNA, responsible for the denaturing of the double helix is explained in terms of the interaction of the bases when they are stacked in the double helical array as delineated by Watson - Crick Model **[49-51].** Semi-quantitative models developed in the 1960s continue to provide significant insight to the melting profiles of DNAs **[44, 52]**. Another major variable generating the melting profiles of the DNA sequences is the choice of the simulation program. In this work we have used the one of the widely used method, MELTSIM (materials and methods). For short oligonucleotides (16-32 bases), Panjkovich and Melo **[53]** performed an extensive comparison of the various methods. In their study it was noted that large and significant differences in the estimations of T_m were obtained while using different methods and no conclusive recommendations were provided on the choice of simulation methods to determine the melting profiles or its accuracy. Here, we are suggesting an approach to classify DNA sequences has potential implication for sequence analysis; DNA sequences could be classified purely from the experimental melting profiles and sequence information is not mandatory as it depends. This method is expected to find wider applications once the sensitivity of the results is established by experiments.

Acknowledgements:
The authors acknowledge Dr. K.W. P. Miller for critical reading of the manuscript. V.V.K. is supported by an NIH grant: Research Infrastructure for Minority Institutions P20MD002732.

References

[1] CJ Benham, SP Mielke, *Annu Rev Biomed Eng* (2005) **7**:21.
[2] JF Marko, ED Siggia, *Science* (1994) **265**:506.
[3] NR Cozzarelli, *Science* (1980) **207**:953.
[4] VA Bloomfield *et al.* (2000) Sausalito, California, *University Science Books*, 794.
[5] CD Bennett *et al. Biotechniques* (2003) **34**:1288.
[6] E Lyon, Expert *Rev Mol Diagn* (2001) **1**:92.
[7] NM von Ahsen *et al. Clin Chem* (1999) **45**:2094.
[8] YJ Chen, X Huang, *Anal Biochem* (2009) **384**:170.
[9] RPalais, CT *Wittwer*, *Methods Enzymol* (2009) **454**:323.
[10] R.Graham *et al*. *Clin Chem* (2005) **51:**1295.
[11] MR Liew *et al. Clin Chem* (2004) **50**:1156.
[12] RA Palais *et al. Anal Biochem* (2005) **346:**167.
[13] CT Wittwer *et al*. *Clin Chem* (2003) **49**:853
[14] CN Gundry *et al*. *Clin Chem* (2003) **49**:396.
[15] C Willmore *et al*. *Am J Clin Path* (2004) **122**:206.
[16] SF Dobrowolski *et al. Hum Mutat* (2005) **25:**306.
[17] J Montgomery *et al. Nat Protoc* (2007) **2**:59.
[18] L Zhou *et al. Clin Chem* (2005) **51**:1770.
[19] L Zhou *et al. Tissue Antigens* (2004) **64:**156.
[20] PN Borer *et al. J Mol Biol* (1974) **86**:843.
[21] KJ Breslauer *et al. Proc Natl Acad Sci* USA (1986) **83**:3746.
[22] RA Dimitrov, M Zuker, *Biophys J* (2004) **87**:215.
[23] J SantaLucia Jr, *Proc Natl Acad Sci* U S A, (1998) **95**:1460.
[24] H SantaLuciaJr *et al*. *Biochemistry* (1996) **35**:3555.
[25] N SugimotoS *et al. Nucleic Acids Res* (1996) **24**:4501.
[26] CM King, AC Wilson, *Science* (1975) **188**:107.
[27] FM Catzeflis *et al. Mol Biol Evol* (1987) **4**:242.
[28] CW Schmid, J Marks, *J Mol Evo* (1990) **30**:237.
[29] CG Sibley, JE Ahlquist, *J Mol Evol* (1987) **26**:99.
[30] SP H.Tanaka *et al*. *J Phys Chem* B (2008) **112:**16788.
[31] H Gille, W Messer *EMBO J* (1991) **10**: 1579.
[32] PA He, J Wang, *J Chem Inf Comput Sci*, (2002) **42**:1080.
[33] K Takahashi *et al. J Hered* (2000) **91**:198.
[34] RD Blake *et al. Bioinformatics* (1999) **15**:370.
[35] RD Volker *et al. Biopolymers* (1999) **50**:303.
[36] Mathworks Inc, Matlab, (2005), *Mathworks Inc., Natick, MA* 01760, USA.
[37] RC Gentleman, *J. Comp Graphics*, (1996) **5**: 299-314.
[38] RC Gentleman *et al. Genome Biol* (2004) **5**: R80.
[39] KTamura *et al*. *Mol Biol Evol* (2007) **24**: 1596.
[40] F Chevenet *et al*. *BMC Bioinformatics* (2006) **7**: 439.
[41] YZ Chen, EW Prohofsky, *Eur Biophys J* (1996) **24**: 203.

[42] O Gotoh, *et al. Biopolymers* (1976) **15**: 655.
[43] M Mandel, *J Marmur* (1968) **12**: 198-206.
[44] C Schildkraut, *Biopolymers* (1965) **3**: 195.
[45] F Liu *et al*. *PLoS Comput Biol* (2007) **3:** e93.
[46] R. Koehler, T N Peyret, *Comput Biol Chem,* (2005) **29**: 393.
[47] DD Long *et al. Biophys Chem* (2004) **110**: 25.
[48] S Payungporn et al. *J Virol Methods* (2004) **120**: 131.
[49] H Devoe, I Tinoco Jr., *J Mol Biol,* (1962) **4**: 500.
[50] H Devoe, I. Tinoco Jr., *J Mol Biol,* (1962) **4**: 518.
[51] I Tinoco, Jr *J Am Chem Soc* (1960) **8**: 4785.
[52] L Kotin, *J Mol Biol* (1963) **7**:309.
[53] A Panjkovichand F Melo, *Bioinformatics,* (2005) **21**. 711.

Edited by P. Kangueane
Citation: Reese *et al.* Bioinformation 4(10): 468-467 (2010)

Supplementary material:

Table 1: Coding Sequences of the First Exon of β-Globin Gene of Eleven Different Species

#	Species	Total	T	C	A	G	GC	Tm	Sequence
			(%)	(%)	(%)	(%)	(%)	(°C)	
1	Bovine	86	20.9	18.6	19.8	40.7	59.3	87.5	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
2	Chimpanzee	105	22.9	19	19	39	58	87.2	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG
3	Gallus	92	16.3	26.1	20.7	37	63.1	88.8	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
4	Goat	86	19.8	19.8	19.8	40.7	60.5	87.9	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
5	Gorilla	93	21.5	20.4	18.3	39.8	60.2	87.7	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
6	Human	92	22.8	20.7	18.5	38	58.7	87.3	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
7	Lemur	92	25	16.3	20.7	38	54.3	85.2	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
8	Mouse	94	24.5	21.3	18.1	36.2	57.5	86.5	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
9	*Opossum*	*92*	*23.9*	*21.7*	*22.8*	*31.5*	*53.2*	*84.9*	*ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG*
10	Rabbit	90	22.2	17.8	18.9	41.1	58.9	86.7	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
11	Rat	92	22.8	19.6	21.7	35.9	55.5	86.9	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG

List of Protein Accession numbers for class II α-chains of MHC proteins

>HumanDRA_M60334.mhctab2seq; >MacacaDRA_L27739.mhctab2seq; >CowDRA_D37956.mhctab2seq; >SheepDRA_M73983.mhctab2seq; >DogDRA_L37332.mhctab2seq; >RatEA_Y00480.mhctab2seq; >PigDRA_M93028.mhctab2seq; >HorseDRA_M60100.mhctab2seq; >RabbitDRA_M28161.mhctab2seq; >WallabyDRA_U18109.mhctab2seq; >HumanDQA_M26041.mhctab2seq; >CowDQA_D50045.mhctab2seq; >SheepDQA_M93430.mhctab2seq; >DogDQA_U42407.mhctab2seq; >MouseAA_M21931.mhctab2seq; >RatBA_X14879.mhctab2seq; >PigDQA_M29938.mhctab2seq; >HorseDQA_L33909.mhctab2seq; >RabbitDQA_M15557.mhctab2seq; >HumanDPA_M27487.mhctab2seq; >RatPA_S80415.mhctab2seq ;>RabbitDPA_M22640.mhctab2seq;
>HumanDNA_M26039.mhctab2seq ;>SheepDNA_Z29533.mhctab2seq ;>MouseNA_M95514.mhctab2seq ;>WallabyDNA_U18110.mhctab2seq ;>ZfishA_L19445.mhctab2seq ;>ZfishA_L19446.mhctab2seq
>ZfishA_L19450.mhctab2seq ;>ZfishA_L19451.mhctab2seq ;>SharkA_M89950.mhctab2seq