Biased amino acid composition in warm-blooded animals

Guang-Zhong Wang and Martin J. Lercher

Bioinformatics group, Heinrich-Heine-University, Düsseldorf, Germany

Among eubacteria and archeabacteria, amino acid composition is correlated with habitat temperatures. In particular, species living at high temperatures have proteins enriched in the amino acids E-R-K and depleted in D-N-Q-T-S-H-A. Here, we show that this bias is a proteome-wide effect in prokaryotes, and that the same trend is observed in fully sequenced mammals and chicken compared to cold-blooded vertebrates (Reptilia, Amphibia and fish). Thus, warm-blooded vertebrates likely experienced genome-wide weak positive selection on amino acid composition to increase protein thermostability.

Corresponding author: Martin Lercher (lercher@cs.uni-duesseldorf.de)

Introduction

Evolutionary molecular biology is mostly concerned with the forces affecting individual genes. However, the observation of variable proportions of Guanine and Cytosine (GC) in different species and in different genomic regions of vertebrates (reviewed in Refs.) have prompted the analysis of forces that may affect the evolution of complete genomes. Against initial expectations, there seems to be no direct relationship between the GC content of prokaryotic protein-coding genes and optimal growth temperature . Similarly, in the case of vertebrates, it was argued convincingly that the 'isochore' structure of high- and low-GC regions is not due to selection, but reflects varying fixation bias of GC over AT in the presence of recombination .

A clear picture emerges only in the study of structured RNAs. The ribosomal RNAs and transfer RNAs of prokaryotes living at high temperatures contain a much larger GC-fraction in their stem regions compared to prokaryotes living at more moderate temperatures , likely because G-C pairs (with three hydrogen bonds) are more stable to thermal fluctuations than A-T pairs (with only two hydrogen bonds). A similar effect is seen in vertebrates: the ribosomal RNA of warm-blooded animals has a higher GC-content compared to that of cold-blooded vertebrates . Thus, in organisms living at elevated temperatures, RNAs that require a specific three-dimensional structure to perform their function appear to be under selection for increased thermostability.

Because proteins also need to retain their three-dimensional structures in the presence of thermal fluctuations, it appears likely that the proteins of thermophile organisms show corresponding signs of thermal adaptation compared to organisms living at lower temperatures. This is indeed observed for prokaryotes: amino acids that according to biophysical considerations tend to lead to stronger electrostatic interactions in protein surfaces are indeed enriched in a set of proteins from thermophiles, while certain amino acids that tend to de-stabilise proteins are depleted . Below, we show that this effect operates on a genome-wide scale in prokaryotes.

The proteins of mammals and birds operate at a constant temperature of 35-42^{o} Celsius, significantly higher than the average temperature in fish or reptiles. Thus, it appears likely that the same trends observed in prokaryotes may also operate on vertebrate proteins: that compared to cold-blooded vertebrates, warm-blooded animals may have proteins with an amino acid composition similarly biased as in prokaryotes living a elevated temperatures. Physiological constraints on multi-cellular animals mean that they cannot live at the temperatures in which thermophilic prokaryotes thrive, and thus we expect their amino acid compositions to be less biased; however, it is conceivable that the same amino acids that stabilize prokaryotic proteins in thermophiles are enriched in warm-blooded relative to cold-blooded animals.

Here, we test this prediction for the first time, using data from 5 fully sequenced warm-blooded and from 6 fully sequenced cold-blooded vertebrates. After demonstrating that the *ERK* measure of biased amino acid composition shows a good

correlation with optimal growth temperature when applied to genome-scale prokaryotic data, we show that the same measure indicates a weak but statistically significant adaptation of amino acid composition to elevated body temperature in warm-blooded vertebrates.

Results

Genome-wide bias in amino acid composition of thermophilic prokaryotes

Based on careful structural alignments of 373 proteins, Glyakina *et al.* showed that among the external residues of proteins from thermophilic prokaryotes, three amino acids (E, R and K) are enriched while seven amino acids (D, N, Q, T, S, H and A) are depleted compared to mesophilic prokaryotes. Thus, the combined proportion *ERK*= E+R+K-D-N-Q-T-S-H-A (where each letter denotes the fraction of the respective amino acid, added for the enriched and subtracted for the depleted amino acids) is elevated for the exterior regions of proteins from thermophiles compared to mesophiles . It has not yet been tested if this measure, which was developed from structural comparisons, is correlated with the optimal growth (or habitat) temperature of individual species when applied to complete protein sequence alignments.

To test this, we use a large set of whole genome sequence data and optimal growth temperature (OGT, ranging from 8°C to 100°C) . This data set contains 204 genomes (180 bacteria and 24 archea), of which 16 are hyperthermophiles (optimal growth temperature OGT 80°C), 11 are thermophiles (OGT=50-80°C), and 177 are

mesophiles (OGT 50°C). In agreement with the earlier observation on a subset of genes , we find a strong correlation between *ERK* and optimal growth temperature (Fig 1, Pearson's R=0.72, $p<10^{-15}$; Spearman's R=0.46, $p=4\times10^{-12}$). This can mostly be attributed to strong differences among hyperthermophiles, thermophiles, and mesophiles ($p=2.1\times10^{-05}$ between hyperthermophiles and thermophiles, p=0.00059 between thermophiles and mesophiles, and $p=4.3\times10^{-11}$ between hyperthermophiles and mesophiles, Wilcoxon rank sum test). However, despite large variation in amino acid composition within mesophiles (Fig. 1), we still see a significant correlation of *ERK* with optimal growth temperature among prokaryotes living at temperatures between 8°C and 50°C (Pearson's R=0.23, p=0.0017; Spearman's R=0.21, p=0.0045).

Organisms living at different ambient temperatures may have different protein repertoires, and thus comparisons of complete genomes are potentially misleading. To circumvent this problem, we performed a complementary analysis restricted to groups of orthologous proteins; orthologs in different species are generally thought to have similar molecular functions. We collected amino acid sequence data from 5 species from hyperthermophiles, 5 species from thermophiles, and 5 species from mesophiles. Using reciprocal best blast hits, we then retained only proteins in each group (hyperthermophiles, thermophiles, mesophiles) that had orthologs in at least one of the other groups (see materials and methods for more details). Comparing the remaining 15293 proteins among groups, it is again clear that $ERK_{\text{hyperthermophiles}} > ERK_{\text{thermophiles}} > ERK_{\text{mesophiles}}$ (Fig 2; hyperthermophiles vs. thermophiles: $p<10^{-15}$,

hyperthermophiles vs. mesophiles: $p<10^{-15}$, thermophiles vs. mesophiles: $p<10^{-15}$, Wilcoxon rank sum tests).

Thus, we confirmed that *ERK* is a useful predictor of temperature adaptation of amino-acid composition on the level of complete proteomes. In the remainder of this paper, we use *ERK* to test for a corresponding effect in vertebrates. To confirm our results, we repeated all analyses in this paper using another predictor from an independent study, the CvP-bias ; all results are qualitatively very similar, supporting the robustness of our conclusions (see Supplementary results).

Biased amino-acid usage in warm-blooded vertebrates

The body temperatures of fish, Amphibia, and Reptilia are closely linked to ambient temperatures, and hence their proteins usually operate at 20-30° Celsius. In contrast, warm-blooded vertebrates (mammals and birds) have a thermoregulation system, which keeps their body temperatures at a constant 35-42° Celsius. Does this difference in temperature result in a discernible selection pressure for increased thermal stability of proteins? If so, we expect to see a difference in amino acid composition between warm-blooded and cold-blooded vertebrates, analogous to the differences reported above for prokaryotic species.

To test this hypothesis, we obtained a total of 526251 protein sequences from 11 completely sequenced species: four mammals (human, *Rattus norvegicus, Mus musculus,* and *Bos taurus*), one bird (*Gallus gallus*), one reptile (*Anolis_carolinensis*),

two amphibia (*Xenopus laevis* and *Xenopus Tropicalis*), and three fish (*Danio rerio, Tetraodon nigroviridis* and *Takifugu Rubripes*). Analysing the amino acid composition of the complete proteomes, we indeed find a small but statistically highly significant shift in *ERK* of warm-blooded compared to cold-blooded verebrates (Fig. 3; $p<10^{-15}$). Again, we confirmed this result by restricting the analysis to orthologous proteins. *Anolis carolinensis* is the closest relative to the warm-blooded animals among the cold-blooded species considered here, and was thus chosen as the reference genome. We identified orthologous proteins in each of the other 10 genomes as reciprocal best blast hits against Anolis. In pair-wise comparisons, all five warm-blooded species show a significantly higher average *ERK* compared to orthologous proteins in *Anolis carolinensis* ($p<0.02$), while this is not the case for any of our amphibia or fish ($p>0.16$; Supplementary Figure 3a-3t). For the 399 orthologs common to all 11 species, *ERK* of the mammal/bird group is significantly higher than for the cold-blooded group (mean *ERK* is -16.136 for the former, while -16.872 for the later, $p=0.0080$, Wilcoxon rank sum test on genomic averages, Table 1).

Elevated *ERK* is not due to biased GC content

The strongest known predictor of amino acid composition at the genomic scale is the GC content of the coding DNA sequences . Thus, it is conceivable that the biased amino acid composition (higher *ERK*) in warm-blooded vertebrates is due to GC

content variation between the genomes of warm-blooded and cold-blooded vertebrates. However, for the 339 co-orthologs studied here, there are no differences in the usage of AT-rich or GC-rich codons between warm-blooded and cold-blooded genomes (p = 0.25 for AT-rich codons and p= 0.13 for GC -rich codons, Wilcoxon rank sum test, Table 1). To further exclude GC content as a confounding factor, we investigated 6227 aligned orthologous coding sequences of human and *Danio rerio* in more detail.

As expected, the human genes encoded proteins with significantly higher *ERK* values than their Danio orthologs (Fig. 4). If these differences in *ERK* could be fully explained by variation in GC content, we would not expect to see different *ERK* values if we restrict our analysis to those aligned codon positions that have the same GC content in human and Danio. Contrary to this expectation, we still see higher *ERK* in the human sequences on these GC-neutral codon positions (Fig 5). Thus, the differences in amino acid composition cannot be attributed to differences in GC content.

Chicken have elevated *ERK* compared to reptiles

Of all cold-blooded animal classes, reptiles – which are paraphyletic due to the exclusion of birds – are the closest living relatives to warm-blooded vertebrates. Thus, we wanted to confirm that the elevated *ERK* values are indeed restricted to warm-blooded animals, by comparing the chicken genome to several hundred recently

published protein segments of three reptilia . Based on best blast hits of the segments against the chicken genome, we constructed 508 protein segment alignments between *Alligator mississippiensis* and chicken, 429 segment alignments between *Chrysemys picta* and chicken, and 138 segment alignments between *Anolis smaragdinus* and chicken (Sup 3). *ERK* in chicken protein segments is significantly higher than in each of the three reptilia species (p=3.89E-16 for Gator and chicken, p=0.00027 for Anolis and chicken and p=0.011 for turtle and chicken).

Discussion

Understanding the factors that influences vertebrate genome composition is of significant important to evolution. Building upon earlier results on 373 aligned structures of prokaryotic protein pairs , we show that genome-wide amino acid composition bias (measured by *ERK*) correlates strongly with the optimal growth temperature of bacteria.

Applying this methodology to 11 vertebrate species, we show that mammalian and bird proteomes have a higher *ERK* value compared to fish, Amphibia and Reptilia. This biased amino acid composition appears not to be due to biases in GC content. As higher *ERK* values are associated with increased thermostability , our findings would be consistent with increased selection for stability against thermal fluctuations in warm-blooded vertebrates. This indicates genome-wide positive selection on amino-acid composition during the change from cold-blooded to warm-

blooded life styles in vertebrates, similar to sequence-based adaptation of microorganisms that switch from being mesophile to being thermophile .

Materials and Methods

Data sources

The genomes of prokaryotic species were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Optimal growth temperatures were taken from Refs. and , except for Chloroflexus aurantiacus J-10-fl, which was obtained from http://genome.jgi-psf.org/finished_microbes/chlau/chlau.home.html. Genome sequences for *Bos taurus*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Mis musculus* and *Rattus norvegicus* were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/) and ENSEMBL (www.**ensembl**.org/info/data/**ftp**/). Protein sequences of *Anolis carolinensis* were downloaded from http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/index.html. Three sets of non-avian reptile protein coding sequences were taken from Ref. .

Calculation of amino acid compositional bias in prokaryotes

The amino acid compositions of the 204 prokaryotic genomes were calculated. For each genome, we then calculated *ERK*=E+R+K-D-N-Q-T-S-H-A (where each capital letter stands for the proportion of this amino acid in the complete genome) , as well as CvP-bias . To identify ortholog groups of proteins with members in

hyperthermophiles (optimal growth temperature OGT 80°C), thermophiles (OGT=50-80 °C), and mesophiles (OGT 50°C), we chose five species from each group. For these 15 species, we performed an all-against-all protein blast search, identifying pair-wise orthologs through reciprocal best blast hits. We then excluded all proteins that hat no orthologs outside their group (hyperthermophiles, thermophiles, mesophiles). All remaining proteins had at least one ortholog outside their group, and where hence retained for our comparison. We then compared mean *ERK* values between groups using Wilcoxon rank sum tests.

Calculation of amino acid compositional bias in 11 vertebrates

For comparison of the 11 vertebrate species, we chose a reptile (*Anolis corolinensis*) as the reference genome. We first identified orthologous protein pairs between Anolis and each of the other 10 veretebrates through a search for reciprocal best blast hits. If an *Anolis carolinensis* protein has an ortholog in each of the other ten genomes, we also included this protein in our co-ortholog list, resulting in 339 groups of ubiquitous orthologs.

GC content as a confounding factor

To check if GC content variation could explain the higher *ERK* values in mammals and chicken, we used reciprocal best blast hits to identify orthologs between human and *Danio rerio*. To align orthologous DNA sequences, we first aligned the translated protein sequences using MUSCLE [] with default settings, and then

replaced the aligned amino acids with their encoding codons. To exclude the influence of differences in GC content, we then re-calculated *ERK* for only those aligned codons that had the same GC fraction in both species.

Comparison of Chicken with three non-avian reptiles

Starting from protein fragments from three non-avian reptile species, we identified putative orthologs in the chicken genome based on the best protein blast hit. Only protein parts that aligned well with the chicken sequence (protein blast e-value < 10^{-5}) were considered further. As before, *ERK* and CvP-bias were calculated for each protein fragment.

References

Figure Legends

Fig 1. Correlation between a measure of amino acid bias related to protein stability, *ERK*=E+R+K-D-N-Q-T-S-H-A, and optimal growth temperature in 204 prokaryotes (Pearson's R=0.72, $p< 10^{-15}$; and Spearman's R=0.46, p=4.008e-12).

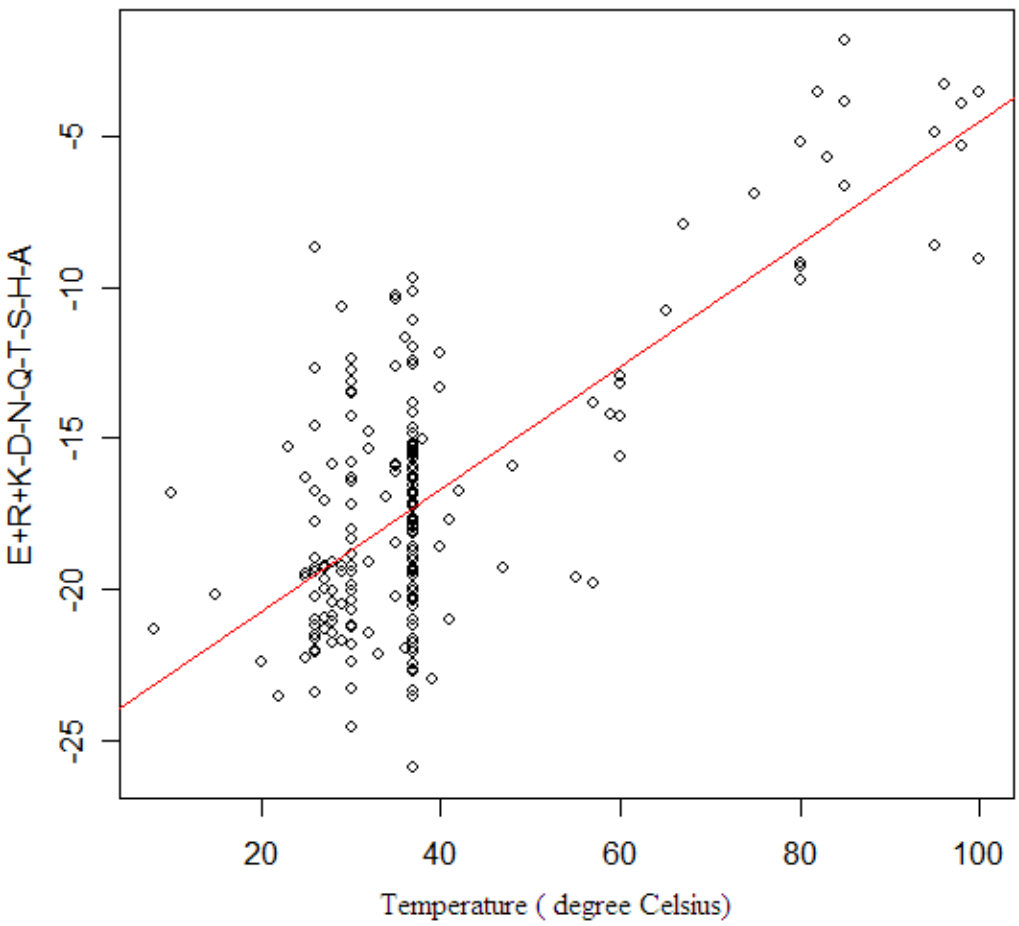


Fig 2. Distribution of amino acid bias, *ERK*, across proteins for 5 species each of hyperthermophile, thermophile, and mesophile prokaryotes.

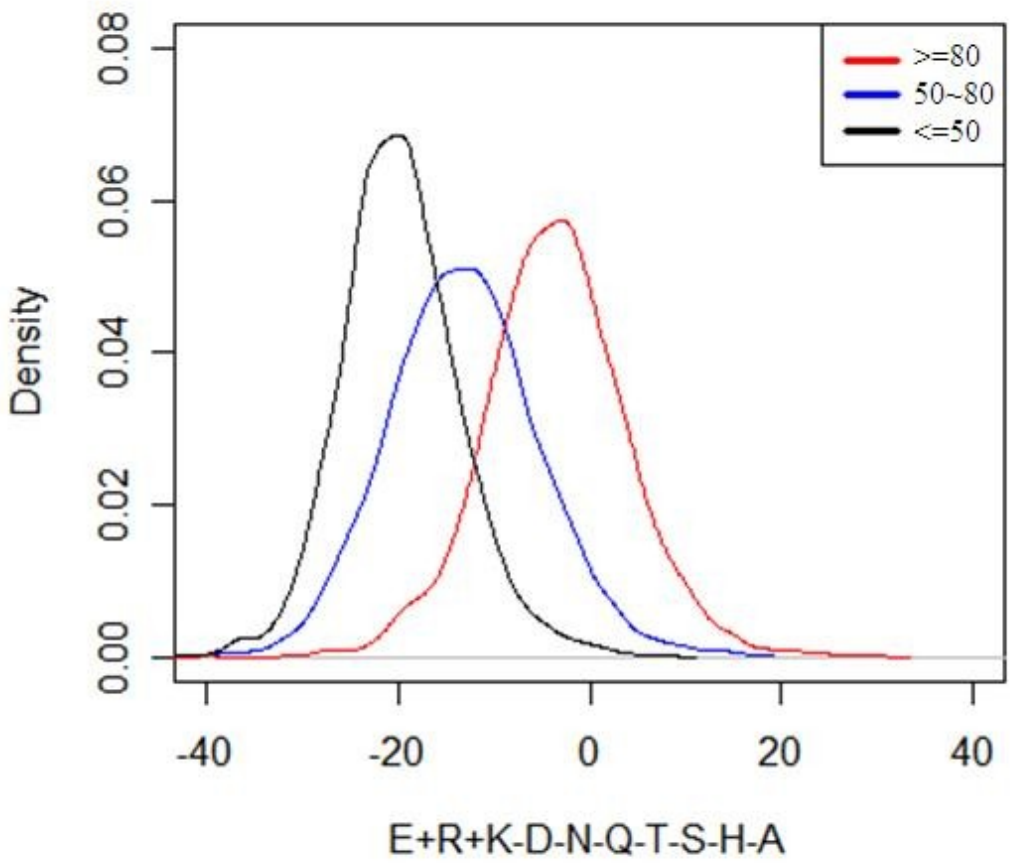


Fig 3. Distribution of amino acid bias, *ERK*, across proteins for warm-blooded vertebrates (mammals, birds) and cold-blooded vertebrates (Reptilia, Amphibia, fish). *ERK* is significantly increased in warm-blooded relative to cold-blooded animals ($p<10^{-15}$).

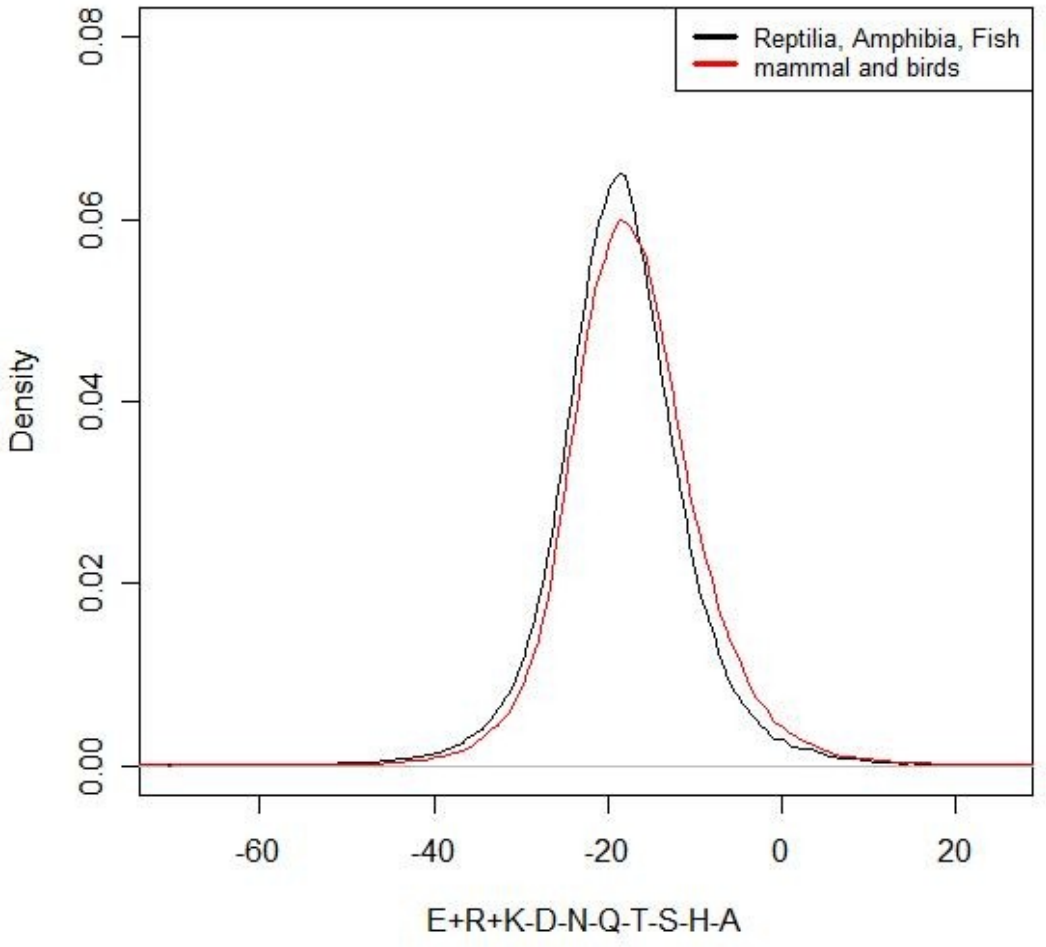


Fig 4. Distribution of amino acid bias, *ERK*, across all codons of orthologous proteins for human and *Danio rerio*. *ERK* is significantly higher in human compared to Danio ($p<10^{-15}$).

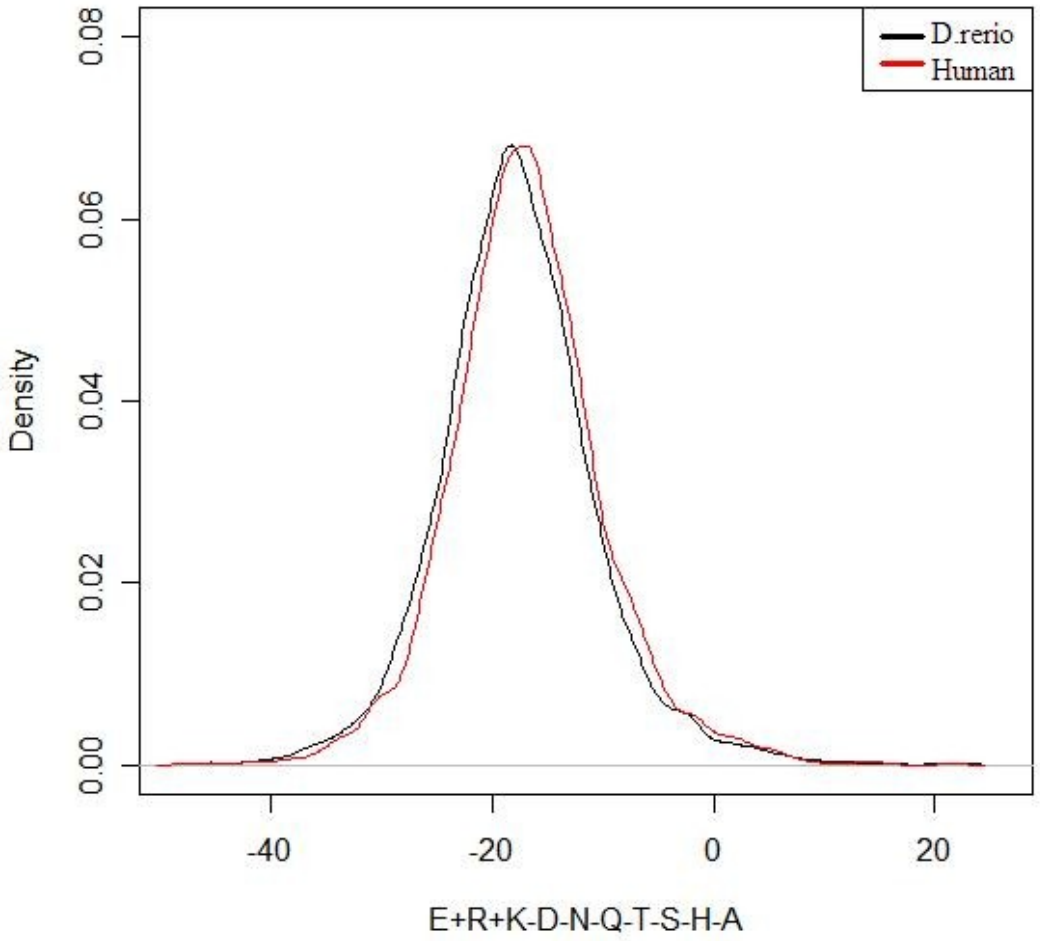
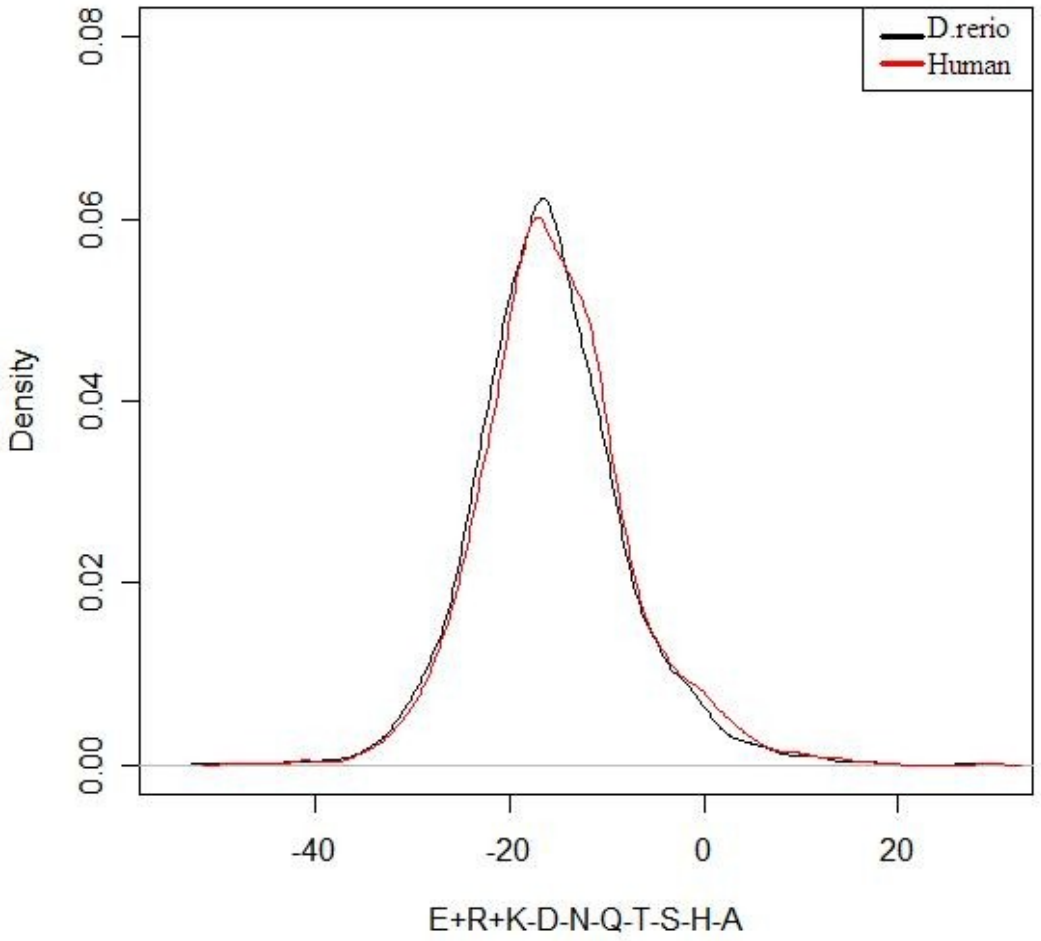


Fig 5. Distribution of amino acid bias, *ERK*, across orthologous proteins for human and *Danio rerio*, using only codons with identical GC content in both species. *ERK* is significantly higher in human compared to Danio ($p<10^{-15}$).



Tables

Table 1. Compositional bias of 339 co-orthologs across 11 vertebrate species. p-values are for comparison of warm- to cold-blooded vertebrates (Wilcoxon rank sum tests).

class	species	ERK	AT-rich	GC-rich
Mammalia	*Mus musculus*	-16.32	23.22	24.56
Mammalia	*Rattus norvegicus*	-16.27	22.82	24.91
Mammalia	human	-16.11	23.82	24.31
Mammalia	*Bos taurus*	-15.97	23.14	25.00
Birds	*Gallus gallus*	-16.01	23.59	24.83
Reptilia	*Anolis carolinensis*	-17.03	23.75	24.45
Amphibia	*Xenopus laevis*	-16.53	25.29	22.52
Amphibia	*Xenopus tropicalis*	-16.53	25.22	22.72
Fish	*Danio rerio*	-16.76	23.85	23.45
Fish	*Tetraodon nigroviridis*	-17.21	22.22	25.45
Fish	*Takifugu rubripes*	-17.17	23.31	23.8
P value		0.0080	0.25	0.13