PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evolution of the isoelectric point of mammalian proteins as a consequence of indels and adaptive evolution

Nicolas Alendé,[1,2] Jens E. Nielsen,[3] Denis C. Shields,[1,2] and Nora Khaldi[1,2]*

[1] UCD Conway Institute of Biomolecular and Biomedical Research, School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Republic of Ireland

[2] UCD Complex and Adaptive Systems Laboratory, University College Dublin, Dublin 4, Republic of Ireland

[3] School of Biomolecular and Biomedical Science, Centre for Synthesis and Chemical Biology, UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

ABSTRACT

Although important shifts in the isoelectric point of prokaryotic proteins, mainly due to adaptation to environmental *pH*, have been widely reported, such studies have not covered mammalian proteins, where *pH* changes may relate to changes in subcellular or tissue compartmentalization. We explored the isoelectric point of the proteome of 13 mammalian species. We detected proteins that have shifted their *pI* the most among 13 mammalian species, and investigated if these differences reflect adaptations of the orthologous proteins to different conditions. We find that proteins exhibiting a high isoelectric point change are enriched in certain GO terms, including immune defense, and mitochondrial proteins. We show that the shift in *pI* between orthologous proteins is not strongly associated with the overall rate of protein evolution, nor with protein length. Our results reveal that insertions/deletions are the main reason behind the shift of *pI*. However, for some proteins we find evidence of selection shifting the *pI* of the protein through amino acid replacement. Finally, we argue that shifts in *pI* might relate to the gain of additional activities, such as new interacting partners, in one ortholog as opposed to the other, and may potentially relate to functional differences between mammals.

Proteins 2011; 79:1635–1648.

Key words: isoelectric point; protein charge; protein evolution; mammal evolution; indels.

INTRODUCTION

The isoelectric point (*pI*) of a protein describes the general electric properties at a given *pH* and corresponds to the value at which the net charge of the protein is zero.

The charge of a protein plays an important role in determining its molecular interactions, and consequently the *pI* value can also be important in this respect, especially when considering differences in *pI* between orthologous proteins, which is the scope of this work. Indeed it has been shown, for example, that the *pI* of interacting proteins in hetero-complexes are usually different,[1] since this yields proteins that are attracted to each other because of their opposite net charges.[1,2] Genome wide searches of the distribution of *pI* values found that there exists a correlation between the *pI* of a protein and its subcellular location.[3,4] The effect is more dramatic for micro-organisms, where it has been shown that the *pI* of membrane proteins follows their ecological niches.[5,6] For example, most membrane protein *pI* is acidic in Escherichia *coli* K12 that live in the basic intestines, and basic in *Helicobacter pylori* that are found in the acidic stomach.[5,7] Moreover, recent work hypothesizes that the eukaryotic cell cytoplasm is *pI* stratified, from *pH* 7.2 at the nuclear membrane, to 6.4 at the cell membrane.[8] This cytoplasmic *pH* stratification hypothesis[8] suggests that observed *pI* changes may have an important impact on signaling, localization of the proteins and as a consequence their reactions and function within the cells of different mammals.

Despite the impact a *pI* change may have on a protein little is known about the variation of *pI* between mammalian orthologs. That said, we would not expect to see the *pI* of orthologous proteins that are conserved between many mammals change dramatically. Indeed the

Additional Supporting Information may be found in the online version of this article.
Abbreviations: *pI*, isoelectric point; indels, insertions-deletions.

Authors' contributions: NA and NK, carried out the analysis; DS and NK supervised, NK conceived the overall idea and strategy, and wrote the paper. DS, NA, and JN contributed to the writing of the paper.
Grant sponsor: Science Foundation Ireland; Grant number: 08/IN.1/B1844; Grant sponsor: Irish Research Council for Science, Engineering & Technology (IRCSET).
*Correspondence to: Nora Khaldi, UCD Conway Institute of Biomolecular and Biomedical Research, School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Republic of Ireland.
E-mail: nora.khaldi@ucd.ie
Received 9 September 2010; Revised 4 January 2011; Accepted 5 January 2011
Published online 18 January 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.22990

habitat conditions, although varying among these mammals, does not reach the marked variation seen among microorganisms. In addition, we expect less intra-organism change between mammals compared to prokaryotes.

Although the net charge and *pI* of a protein may not always give strong insights into its function and functional interactions, it is likely that changes in these values between orthologous proteins maybe frequently associated with changes in interactions and function.

In this work we investigate changes in *pI* between orthologous mammalian proteins in 13 publicly available sequenced genomes (human, cow, chimp, monkey, rat, mouse, dog, cat, opossum, platypus, horse, rabbit, and guinea pig), and examine the reasons for this shift.

METHODS

Data

The human, chimp, monkey, mouse, rat, guinea pig, rabbit, cat, dog, horse, cow, opossum, and platypus protein sequences were downloaded by FTP from the Ensembl database at: http://www.ensembl.org/info/data/ftp/index.html. For each protein, we selected the longer sequence among its different isoforms, or annotations, because it represents the better coverage of the gene's sequence.

Orthologs and sequence evolution

To find orthologs, we identified best mutual way BLASTP hits among human, chimp, monkey, mouse, rat, guinea pig, rabbit, cat, dog, horse, cow, opossum, and platypus. In other words, a protein is considered as an ortholog in all 13 mammals if it satisfies 78 reciprocal best hits (the *E* values we considered in each step are $\leq 10^{-4}$). This method resulted in a set of 1412 putative orthologs among the 13 mammal species. We also used the set of orthologs between human and mouse, represented by the reciprocal best hit between both genomes, and we detect 16527 proteins (*E* value $\leq 10^{-4}$).

Calculating the isoelectric point

The common *in vitro* estimation of *pI* is performed using a bidimensional electrophoresis gel. For the purpose of our genome wide study we estimated the *pI* from the protein sequences, errors introduced in the calculation of *pI* from the sequence mainly arise because proteins can alter the *pKa* values of their constituent titratable groups.[9,10] It is possible to calculate the *pKa* with a good accuracy when a high-resolution 3D structure of the protein is available.[11,12] However, such calculations will not necessarily increase the accuracy of the predicted *pI* values due for example to differences in the protein structure because of crystallization conditions.[13]

To calculate the *pI* we first cleaved off the signal peptide from each protein using a HMM search with SignalP-HMM.[14] The rest of the sequence was incorporated into an in-house perl script for the calculation of the *pI* that uses the Henderson-Hasselbalch equation. The script searched for the number of R, K, Y, C, H, E, and D that are implicated in the *pI* of a protein. Each of the previous amino acids was assigned a *pKa* value, 12.48, 10.54, 10.46, 8.18, 6.04, 4.07, and 3.9, respectively, 8.0 for the N-terminus, and 3.1 for the C-terminus. The charge that is contributed by arginine to the charge on the entire protein is the product of the charge of arginine at a specific *pH* with the count of arginine in the full sequence. We can then calculate an estimated charge for the protein at any particular *pH*. To determine the *pI,* which is the *pH* value at which the estimated charge is zero, we estimated an initial *pH* at which the overall charge of the protein is positive and one where the charge is negative. We then used a bisection method to estimate to a 10^{-2} precision the value that renders the overall charge null. In this method cysteines are assumed not to take part in disulfide bridges, and thus may take on negative charges. The algorithm for *pI* calculation also assumes that residues in the sequence are independent of each other; N- and C-termini have fixed *pKa* values, except modified termini which are ignored; and finally that modified residues, such as phosphorylated ones are ignored. The above is also true for tyrosine (Y). Although a more thorough analysis would exclude known disulphide bridged cysteines from the analysis, this would have introduced a technical bias, because some proteomes have a much higher disulphide annotation quality than others. Accordingly, we opted to include cysteines in the *pI* calculations. In our *pI* calculation we do not take into account the effects of potentially *pI*-changing post-translational modifications to the protein. Phosphorylation is the most common post-translational modification and generally introduces a -2 charge in a protein per phosphate group added. Phosphorylation sites are difficult to predict with high accuracy, and the percent phosphorylation at a site is often hard to determine, although resources such as Phosphobase[15] and NetPhosK[16] have made considerable headway. For the purposes of this study we chose to ignore phosphorylation sites, rather than changing the *pI* by introducing phosphorylation site predictions of unknown accuracy. Similarly, we ignored glycosylation since it is currently not possible to predict the location and effects of such modifications accurately.

Defining significant *pI* shifting proteins

A protein is considered as significantly shifting in, for example in mouse, if the difference between its *pI* and that of its ortholog in human is higher than a threshold that is determined from the differences in *pI* of all orthologs between human and mouse (Table I). Setting a

Table I
Threshold of Significant Shifts in *pI* And Number of *pI* Shifting Proteins

Pairwise comparison		
Human-chimp	0.48339	2007
Human-mouse	1.06935	2019
13 species		
Chimp	0.381	120
Monkey	0.6066	178
Mouse	0.92286	175
Rat	0.96972	199
Guinea pig	1.05468	137
Rabbit	1.5501	160
Cat	0.92286	179
Dog	0.98145	152
Horse	0.91551	182
Cow	0.92286	167
Opossum	1.23192	169
Platypus	1.8384	158

Three values are provided for pairwise and 13 mammalian species comparison, these are the species name, the significant threshold, and the number of proteins we considered in each comparison.

threshold of *pI* between two species is somewhat arbitrary because the data does not follow a known distribution, for this reason we used a nonparametric formula to define the threshold of significance. This threshold is calculated using the median, and third quartile of the absolute shift in *pI* between orthologous proteins this is: threshold $= 2 \times$ (3rd quartile–median).

Go-term enrichment

We used the online tool GOrilla[17,18] to estimate any possible enrichment in the sets of genes that are coding for proteins that present a significant shift in their *pI* and we searched the three elements: process, function, and component. As GOrilla only recognizes the gene ID of a limited number of genomes, we used human as a background set for our analyses, and assumed that the GO-term annotations are shared between orthologous proteins. As the ID of the genes in the background and in the target sets must belong to the same species, we used the IDs of the human genes to represent the sets of significantly shifting proteins in each of the remaining 12 species. The GO-term association *P*-values that are presented in the text and in Table III are corrected for multiple hypothesis testing, and are provided by GOrilla.

Non-synonymous substitution rate (*dN*) calculation

All the orthologous proteins between human and mouse (16527) were aligned using clustalW.[19] We used a maximum likelihood method to calculate the nonsynonymous substitutions based on the alignments. This was performed using codeml from the PAML package,[20] using the default parameters. The results consisted in two *dN* values, which represent the distance of the human and the mouse sequences provided in the alignment to a maximum likelihood predicted ancestral sequence. We used the difference between the human and mouse *dN*, referred to as ΔdN in the text as a measure of difference in the substitution rate of both proteins. We also used ΣdN, which represents the total amount of substitution separating both proteins (ones affecting a protein plus ones affecting its ortholog).

Cellular location

Using Biomart (http://www.ensembl.org/biomart/index.html) we downloaded the cellular component annotations of the proteins from the Ensembl release (version 55). The proteins were separated into different cellular locations (Supporting Information file 2): plasma membrane, cytoplasm, cytoskeleton, lysosome, endosome, mitochondrion, peroxisome, Golgi, reticulum, nucleus, secretory granule, and extracellular. We used synonyms to group similar locations under the same name (Supporting Information file 2). For example, component annotations containing the words tubulin, actin, centrosome, keratin, or microtubule are associated with the cytoskeleton location. Supporting Information file 2 shows the list of all the synonyms used. Indeed, we obtained a list of cellular locations for each protein in human and a list for each of its ortholog in mouse.

When we compared the locations of the orthologs, we defined related locations that we considered as being similar. For example, if one protein is located in the endosome in mouse and its ortholog in human is located in the lysosome, we did not consider it as a difference in location because the endosome is meant to merge with the lysosome after its formation.[21] Two lists of locations are different if there is at least one location in one of those two lists that is not shared by, or that does not have a related location in, the other list. The locations that we considered as being related, and thus, not generating a difference between orthologs are represented on Supporting Information Fig. 3.

Statistical analyses

All the statistical analyses were performed using R (2.8.1). We performed a nonparametric Kruskal-Wallis to test the equality of the median length of the cluster of *pI* shifting proteins with the median length of the whole human proteome. A Chi-square test was used to test for certain categories being enriched. The PCA analyses were performed using the Rcmdr package. We used the *pI* shift, either the *dN* shift or the sum of *dN* and either the length or the length shift as active variables for the analyses.

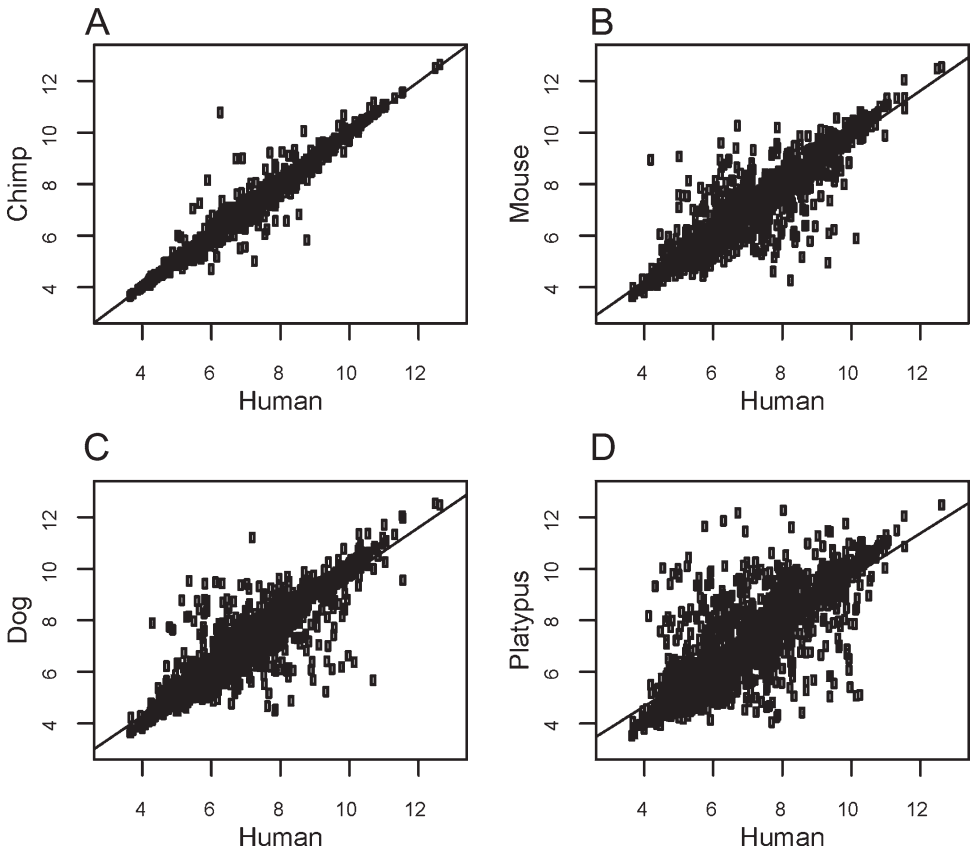

Figure 1

Scatterplot and correlation line of *pI* values of human proteins versus their orthologs in chimp, mouse, dog, and platypus. The black line in each figure represents the regression line. The X axis in all four plots represents human *pI*, and the Y axis represents the *pI* value of the orthologous proteins in the following species: **A**: Chimp with a correlation of 0.99, **B**: Mouse with a correlation of 0.92, **C**: Dog with a correlation of 0.9, and finally, **D**: Platypus with a correlation of 0.78.

RESULTS AND DISCUSSION

We calculated the *pI* of all proteins in all 13 mammals (see methods). We used human as a reference genome, and plotted the *pI* of all 13 species against the human *pI*. We used human as a reference because of the high quality and coverage of its GO-term annotations, which is used further in our analysis. Figure 1 shows that most proteins conserve their *pI* despite the long evolutionary distances, such as that separating human from platypus [Fig. 1(D); mean distance between *pI* in human and mouse = 0.02; standard deviation = 0.86]. However, some proteins have dramatically shifted between human and the 12 mammals. Some have even experienced jumps from an acidic *pI* in the other mammals to basic *pI* in human, and vice versa (821 have experienced this type of shift between human and mouse, where the first *pI* is higher than 7.4, whereas the ortholgous *pI* is lower than 6.6). These large shifts are unexpected given that it is thought that orthologous proteins, especially ones that have been conserved in numerous genomes, have also conserved functions. Changes in Histidine between orthologous proteins may contribute unequally to buffering under physiological conditions (*pH* = 7.4), so changes in *pI* that are also influenced by Histidine may just reflect important changes in Histidine between orthologs. We carried out a search of differences in Histidine counts between orthologous proteins, and found that only 16% of proteins that have experienced shifts in *pI* present changes in Histidine between human and chimp greater or equal to 3 residue differences. In other words, 84% of the entire set of *pI* shifting proteins present 0 to two changes, which will limit buffering affects in most proteins.

Another observation that is noteworthy is that the highest shifts in *pI* affect proteins around *pH* = 7. Relatively speaking, fewer proteins are found in this region of physiological *pH*, however, it seems like these proteins have a more flexible *pI*. A possible explanation may be that their *pI* is not that relevant for some of these proteins, and that the protein can vary the *pI* between different mammals without affecting its activity.

Table II
Go-term Enrichments for the Orthologous Proteins that have Significantly Shifted Between Human-Chimp, and Between Human-Mouse Pairwise Comparisons

Pairwise comparison	Go category	Go-term enrichment	No. proteins	P-*value*
Human-chimp	Process	cellular defense response	14	4.62×10^{-4}
		cell migration[a]	32	3.06×10^{-4}
		purine ribonucleotide metabolic process	19	1.88×10^{-4}
		immune system process[a]	96	4.41×10^{-4}
	Function	nucleobase, nucleoside, nucleotide kinase activity	13	3.34×10^{-4}
Human-mouse	Process	purine ribonucleotide triphosphate metabolic process	15	1.56×10^{-4}
		respiratory electron transport chain	13	3.47×10^{-4}
		immune response[a]	85	4.07×10^{-4}
	Function	hormone activity	28	1.64×10^{-5}
		cytokine activity[a]	34	5.29×10^{-5}
		tumor necrosis factor receptor binding[a]	8	6.97×10^{-4}
		water channel activity	5	8.22×10^{-4}
	Component	membrane	765	0
		integral to membrane	610	0
		extracellular region	245	6.84×10^{-6}
		mitochondrion	146	3.44×10^{-5}
		extracellular space	78	1.52×10^{-4}

[a]Immune related proteins.

We wanted to analyse these sets of proteins with shifting *pI* and investigate if they were enriched in certain GO-terms, also if these changes are a consequence of selection.

Gene Ontology term enrichment

Pairwise comparisons

To investigate if certain types of proteins experience significantly larger changes in *pI*, we searched for GO-term enrichment for the human-chimp and human-mouse comparisons. We considered a shift in *pI* between human and chimp to be high if it was greater than 0.48, and between human and mouse if it was greater than 1.07 (Table I; see methods for rationale behind choice of these cut-offs).

We further tested for the significance of this threshold by randomly assigning *pI* values to proteins, and found that our set thresholds are in all cases significant ($P <= 0.01$). We show that immune related proteins are significantly enriched in the set of proteins that have shifted in *pI* (Table II). Perhaps because they are subject to different substrates, and thus have a freer *pI* shift range, or have adapted to these different substrates, and have changed their *pI* as a consequence. This is also true for the enrichment observed between human and mouse, where the *pI* shifting proteins seem to be enriched in immune related proteins (Table II). Further, the component enrichment in mouse is associated with membrane and extracellular related proteins arguing once more for a possible adaptation to different substrate stimuli. We wished to ensure that our analysis was not biased by the inclusion of transmembrane proteins. Indeed, the estimates of *pI* for transmembrane proteins are not expected to be as accurate as those predicted for soluble proteins, because the nonpolar nature of the lipid membrane will induce fairly large *pKa* shifts in many surface titratable groups. We thus investigated the number of transmembrane proteins in our set of *pI* shifting proteins based on the term membrane in their description (plasma membrane, golgi membrane, mitochondrion membrane). We find that transmembrane proteins represent 35.8% of the human-chimp proteome (and 39% in human-mouse). Similarly, the *pI* shifting proteins contain 34% between human-chimp, and 35.8% in human-mouse comparison. These results indicate that transmembrane proteins are not affected by change in *pI* more than expected. Thus, although the actual *pI* of the transmembrane proteins are slightly artifactual, there is no indication that the changes in *pI* of those proteins are biased as a result.

Proteins conserved in all 13 mammals

Similar to the pairwise comparison, we examined if certain types of proteins that are conserved in all 13 mammals perform certain functions that will require a change in *pI*. We selected 12 sets of proteins, each set represents proteins with significant *pI* shift in one of the 12 species in comparison to human (Table I). We analyzed the results for each pairwise comparison independently to search for GO term enrichment for proteins that have shifted between human and cow for example. Next, we analyzed the overlap between the different pairwise comparisons to detect if there were certain types of GO terms that tend to change their *pI* in multiple species. Table III shows the GO-term enrichment for each pairwise comparison of the conserved orthologs between the

Table III
Go-term Enrichments for the Orthologous Proteins Among the 13 Species

Species	Go category	Go-term enrichment	No. proteins	P-*value*
chimp	Function	transforming growth factor beta receptor, cytoplasmic mediator activity (transmembrane receptor protein serine/threonine kinase signalling protein activity: 2)	2	3.11×10^{-4}
monkey	Process	cellular lipid metabolic process	16	3.66×10^{-4}
		glycyl-tRNA aminoacylation (tRNA metabolic process: 7)[a]	2	3.46×10^{-4}
	Function	glycyl-tRNA ligase activity (catalytic activity: 80)[a]	2	3.46×10^{-4}
	Component	Mitochondrion[b]	22	4.34×10^{-4}
mouse	Process	carbohydrate metabolic process	15	1.64×10^{-4}
		tRNA processing[a]	5	5.73×10^{-4}
	Function	tumor necrosis factor receptor binding	3	9.92×10^{-4}
	Component	Mitochondrion[b]	23	1.19×10^{-4}
rat	Process	carbohydrate metabolic process	15	6.60×10^{-4}
	Function	ethanolaminephosphotransferase activity	2	4.29×10^{-4}
	Component	Mitochondrion[b]	30	6.92×10^{-7}
guinea pig	Process	ncRNA processing[a]	7	2.08×10^{-4}
		ncRNA metabolic process[a]	8	2.37×10^{-4}
	Function	lipoic acid binding	2	4.01×10^{-4}
		S-acyltransferase activity	2	9.91×10^{-4}
	Component	Mitochondrion[b]	19	2.15×10^{-4}
rabbit	Process	coenzyme metabolic process (cofactor metabolic process: 7)	7	2.12×10^{-4}
		cellular lipid metabolic process	15	3.06×10^{-4}
		ncRNA metabolic process[a]	8	6.53×10^{-4}
	Function	translation initiator factor activity (translation factor activity, nucleic acid binding: 6 | translation regulator activity: 6)	5	1.51×10^{-4}
		lipoic acid binding	2	5.41×10^{-4}
	Component	Mitochondrion[b]	21	2.12×10^{-4}
		peroxisomal matrix (peroxisomal part: 4 | microbody part: 4)	2	5.41×10^{-4}
cat	Process	tRNA processing (tRNA metabolic process: 9 | ncRNA metabolic process: 13 | ncRNA processing: 11 | RNA processing: 15 | RNA metabolic process: 20)[a]	9	4.42×10^{-6}
		cellular respiration	3	6.48×10^{-4}
		cofactor biosynthetic process	6	2.10×10^{-4}
	Function	lipoic acid binding	2	7.03×10^{-4}
	Component	Mitochondrion[b]	24	7.44×10^{-5}
dog	Process	tRNA processing (tRNA metabolic process: 6)[a]	5	2.89×10^{-4}
	Component	Mitochondrion[b]	21	1.00×10^{-4}
		peroxisomal matrix	2	4.88×10^{-4}
horse	Function	cytokine activity	8	2.42×10^{-4}
cow	Process	glutamine metabolic process	3	7.21×10^{-4}
	Function	methyltransferase activity (transferase activity, transferring one-carbon group: 9)	9	2.77×10^{-5}
	Component	mitochondrion (cytoplasmic part: 58)[b]	25	5.66×10^{-6}
opossum	Process	tRNA processing (tRNA metabolic process: 7)[a]	6	5.06×10^{-5}
		ganglioside metabolic process (glycosphingolipid metabolic process: 3)	3	5.83×10^{-5}
	Function	tRNA binding[a]	3	7.87×10^{-4}
	Component	Mitochondrion[b]	22	2.25×10^{-4}
platypus	Process	ncRNA processing[a]	7	4.63×10^{-4}
		base-excision repair	3	6.00×10^{-4}
		sex differentiation	3	6.00×10^{-4}
		vitamin metabolic process	5	7.45×10^{-4}
		protein folding	7	8.58×10^{-4}
	Function	uracil DNA N-glycosylase activity (hydrolase activity, hydrolizing N-*glycosyl compound: 3)*	2	5.21×10^{-4}

[a]Noncoding RNA (ncRNA) related proteins
[b]Mitochondrial proteins.

13 mammals. The results are different to the pairwise comparisons in that there is no over-representation of immune related proteins. This is because immune related proteins are not well conserved when we consider conservation in 13 species that span mammals, and are thus not present in this set. This table shows that the proteins that change significantly in *pI* between species present significant enrichment in many functions and processes. And in 9 out of the 12 comparisons, the sets of proteins are significantly associated with mitochondria. The pro-

teins involved in the enrichments in mitochondrial components for all those 9 species make up to 73 different proteins among which 19 (26%) are involved in the respiration and 14 (19%) in protein expression (ribosomal proteins, ncRNA related proteins including: tRNA synthetase and tRNA transferase). The other 55% of proteins did not cover a common group. It is not clear why such a large number of proteins presenting significant change in *pI* are associated with mitochondria. We find that only 17.8% (13 out of 73 proteins) of the mitochondrial shifting proteins between human and mouse comparison are transmembrane proteins, indicating that this is not the main reason behind their shift in *pI*. Possible interspecies differences in intracellular *pH* in the mitochondrion cannot explain this trend, because if this were the case we would expect that such a scenario would be accompanied by a directionality, which we did not observe. For example, in the mouse-human comparison we have 9 proteins that have reduced their *pI*, while 10 show an increase (Supporting Information Fig. 1). This difference does not yield a significant directional trend to the mitochondrial-proteins *pI* shift. We find similar results with other species comparison. We do note, however (Supporting Information Fig. 1), that the *pI* shifts are confined to the proteins with *pI* in the range 6–10; mitochondrial proteins with more extreme predicted *pI* values remain constant. This might be due to the possibility that fewer amino acid substitutions are required to shift a central *pI* relative to a more extreme *pI*. Other species-specific *pI* differences (Table III) are not easily related to specific phenotypic or genotypic differences. For platypus, Table III shows that a subgroup of the proteins that present the highest shift in their *pI* to human orthologs are implicated in sex differentiation (3 proteins: FKBP4, DMRT2, and RQCD1; $P = 6 \times 10^{-4}$). This observation may relate to the complex sex determination pathway in this species.[22]

Are there certain GO terms that tend to shift their pI in multiple species?

Table III also shows the relatively small number of processes and functions that overlap over 2 or more species. Indeed, only three processes and one function are found in common between two or more species when compared with human, these include the two processes 'lipid metabolic process' that is found in monkey and rabbit, and 'carbohydrate metabolic process' which is found in mouse and rat (Table III). Also, the function 'lipoic acid binding' that is found in guinea pig, rabbit, and cat. Only one process, which includes noncoding RNA processing terms, is present in 8 species out of the 12 (human is a reference), this includes the terms 'tRNA' and 'ncRNA'. Indeed, enrichments in tRNA processing are found in cat, dog, mouse and opossum; ncRNA processing in guinea pig, platypus and rabbit, and finally glycyl-tRNA aminoacylation in monkey (Table III). A possible explanation for this enrichment in non-coding RNA is a regulation of gene expression level. Indeed, it is accepted that the great phenotypic differences between mammalian cells of different types are to a large extent determined by differences in gene expression patterns.[23] Although transcriptional control is the main level at which gene activity can be regulated, it is by no means the only mechanisms and regulation also occurs by RNA processing.[24–28] This significant result might imply that the differences of *pI* in the proteins implicated in non-coding RNA processing underlies differences in regulation of gene expression and thus greater phenotypic differences between mammalian genomes.

What is causing the shift in pI?

The shift in *pI* is neither strongly associated with the overall rate of protein evolution nor with the protein length.

Because mutation into and out of R, K, Y, C, H, E, and D is one of the causes of shifts in *pI*, we would expect that proteins with a fast rate of evolution (ones with a high nonsynonymous substitution rate *dN*) would have more *pI* shifts, if the *pI* were not under purifying selection. We make this assumption, because more rapidly evolving proteins absorb more changes in a given evolutionary window, and that flexibility in terms of structure gives them more opportunity to evolve more markedly diverged *pI*s. At an extreme, a completely conserved protein will only change by one amino acid in a long evolutionary period, during which selective processes on *pI* may have more time to act. To investigate this, we calculated the nonsynonymous rate of evolution *dN* of human-mouse orthologs, and determined a value ΔdN that represents the absolute value of the difference in the rate of protein evolution between each species for each ortholog. We used a principal component analysis method presented in Figure 2 to determine the relation between the three variables: ΔpI that represents the difference in *pI* between the human-mouse orthologs; the length of the proteins in human; and the evolutionary difference between both orthologs (ΔdN). We used the principal component for both the set of all orthologous proteins between human and mouse, and the set of *pI* shifting proteins we selected. Two parameters are dependent in a PCA if they both can be simultaneously projected on at least one axis; the length of the projection on the axis is important and is representative of the strength of their dependence. Figure 2 shows that the sequence length and ΔdN are dependent on each other. Indeed, both can be simultaneously projected on all three axis 1, 2, and 3. Inversely, neither the sequence length nor ΔdN depend on ΔpI (and vice versa). Indeed there is no clear projection of ΔdN and ΔpI on an axis, nor that of the sequence length and ΔpI. The results show that

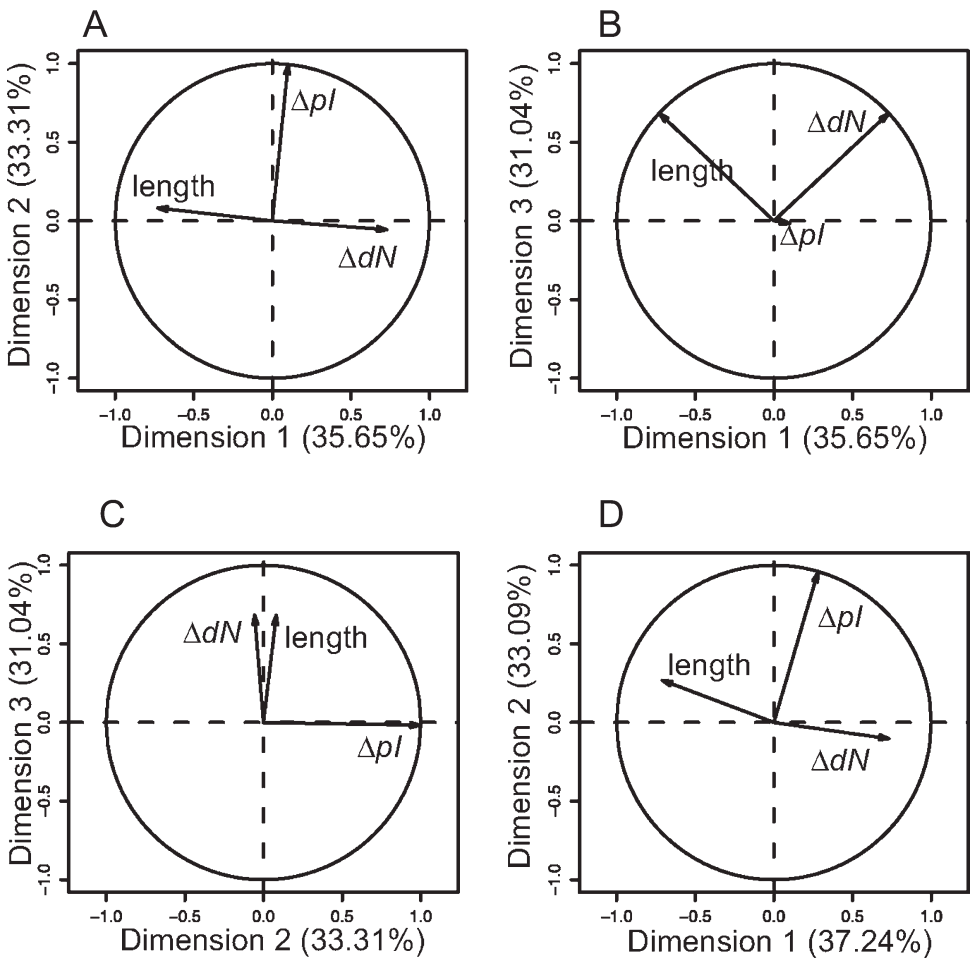

Figure 2

PCA analysis of the relationship between the shift in *pI* (Δ*pI*), the sequence divergence (Δ*dN*) and the length of the proteins between human and mouse. **A**: Presents the contribution of the three variables to dimensions 1 and 2 for all the orthologous proteins between human and mouse. **B**: Similar to (A) but for dimensions 1 and 3. **C**: Similar to (A) but for dimensions 2 and 3. **D**: Presents the contribution of the three variables to dimensions 1 and 2 for the *pI* shifting proteins between human and mouse.

Δ*pI* for both sets is independent of both Δ*dN* and length (see Fig. 2). Indeed, the contribution of Δ*pI* to axis 2 is 99% for all the orthologs [Fig. 2(A)] and 90% for the set of *pI* shifting proteins [Fig. 2(D)], this axis is independent from axis 1 and 3 that are represented by a combination of length and Δ*dN* [Fig 2(A–C)]. We further produced a multiple regression with the human-mouse orthologs and found that the fraction of variation in Δ*pI* that is explained by both variable Δ*dN* and length is only 1.7% (adjusted $R^2 = 0.017$). The high *pI* differences that are observed are not simply a consequence of accelerated sequence evolution. On the other hand, the results tells us that although a protein can diverge in sequence it can at the same time maintain its *pI*, arguing for a selection maintaining the *pI* regardless of the substitution rates.

Indels as a main cause for pI shift

If the significant shifts observed in the *pI* between proteins were due to indels, we would expect to see a correlation between the two parameters, the *pI* shift, and the differences in length between orthologs. If, however, it were due to substitution, we should not observe a correlation. To investigate this question we observed the relation of the two previous parameters using a principal component analysis method. A closer look at the PCA on Figure 3(A-C) shows that *pI* and Δlength are dependent. Indeed both can be projected on axis 1 [Fig. 3(A)], axis 2 [Fig. 3(B)], and axis 3 [Fig. 3(A)]. The *pI* is not independent of length, they seem to contribute to each dimension equally [Fig. 3(A-C)]. They can be positively correlated (dimension 1), and negatively correlated on (dimension 2, and 3).

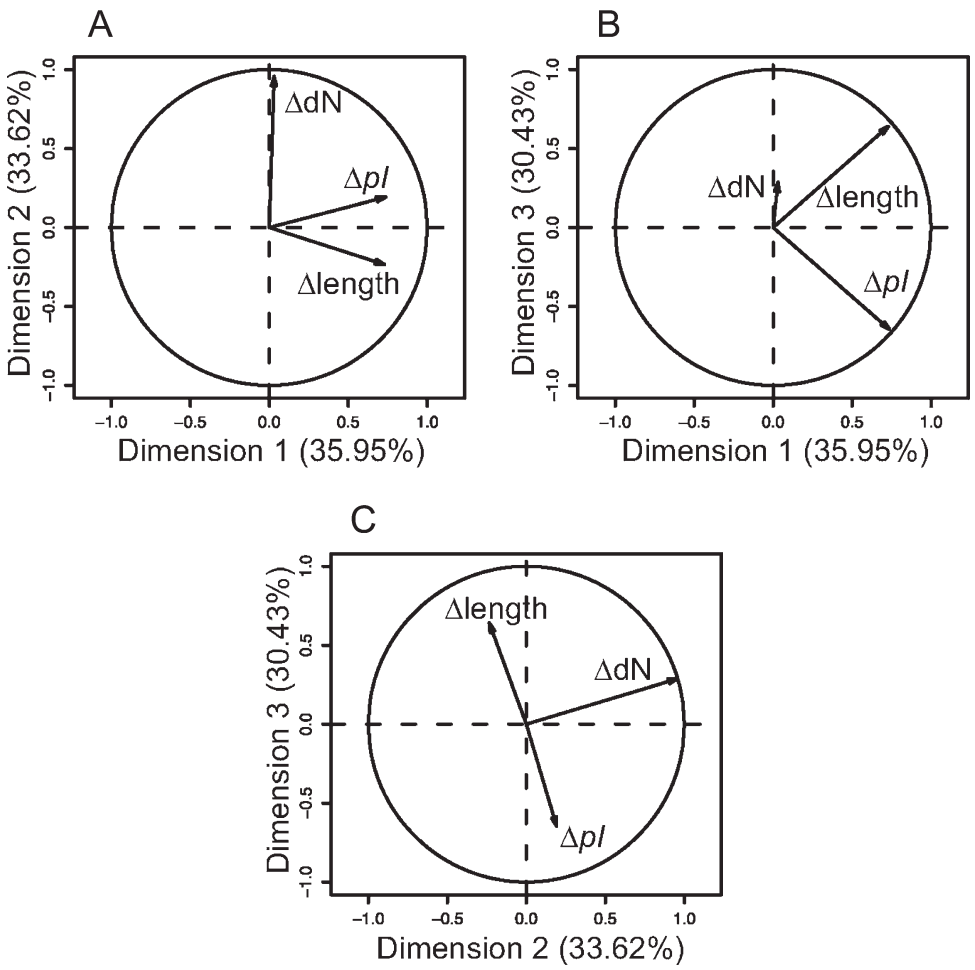

Figure 3

PCA analysis of the relationship between the shifts in *pI* (Δ*pI*), the sequence divergence (Δ*dN*), and the difference in length of the proteins (Δ*length*) between human and mouse. **A**: Presents the contribution of the three variables to dimensions 1 and 2 for the *pI* shifting proteins between human and mouse. **B**: Similar to (A) but for dimensions 1 and 3. **C**: Similar to (A) but for dimensions 2 and 3. **D**: Presents for the *pI* shifting proteins between human and mouse Δ*pI* versus Δ*length* (Δlength $\leq$200 amino acids).

We find that 96% of the proteins that have significantly shifted in *pI* in human-mouse comparison have also experienced some form of indel of 1 or more residues [Fig. 3(D)]. This percentage is reduced to 54% when we consider proteins that have experienced a loss or gain of 10 residues or more. This result argues that more proteins that have significantly shifted in *pI* have also experienced important indels, arguing that indels are a major cause for shifting the *pI* of proteins between species. However, is this result obvious? Do we always (or at least in the majority of cases) observe indels accompanied with shifts in *pI*? Given that approximately a third of the amino acids of globular proteins of average size are charged, one would expect that the majority of proteins that experience indels also consequently experience a shift in *pI*. To examine if this is the case we first calculated the average indel length across the *pI* shifting proteins, which we found to be 10.7% of the length of a protein. Of the 1591 proteins that have experienced such indels, only 736 have shifted in their *pI*, while a bigger number 855 have not shifted their *pI*. This might indicates that most indels are retained in the protein only because they do not affect its *pI* arguing once more for selection on retaining the *pI* value of a protein. A multiple regression of the *pI* shifting proteins between human-mouse shows that there is a positive and significant correlation between Δ*pI* and changes in the protein's length ($P = 1.27$ e-05), and that this variable accounts for 2.5% of the variation in Δ*pI* (adjusted $R^2 = 0.025$). Although both *P*-values for the sequence length in Figure 2, and indel length in Figure 3 seem to be close (1.7, and 2.5% respectively) the PCA shows clearly that only indels (Δlength) explain the shift

in ΔpI. The small proportion of the variability accounted for by the regression in the data indicates that shifts in *pI* are not strongly dominated by such factors, and that the great majority of the variance may reflect other processes of which presumably a subset represent adaptive changes in the *pI*.

Studying the extent of change in *pI* in relation to insertions and deletions in protein evolution is complex, because the regions in proteins that are most prone to insertions and deletions are also more prone to protein disorder. Because many disordered proteins have a preponderance of charged amino acids (typically occurring in short low-complexity runs of residues of similar charge), single deletion or insertion events in disordered regions can have a marked influence on the *pI* of the protein, compared with what you might expect if amino acid composition and protein structure were random. A complete evolutionary model of charge change in proteins would need to model carefully the different "flavors" of disordered sequences,[29] as well as the rate of different classes of indel mutational events in different protein regions.

Selection can explain part of the pI shift

Two results support selection as opposed to neutral evolution of *pI*. Under neutral selection on *pI*, we would expect to observe shifts in *pI* mainly in proteins with a high substitution rate. However, as mentioned above we did not detect any correlation between these two parameters and both seem for most of the *pI* shifting proteins independent of each other [Fig. 2(A–C)]. In addition, as discussed above, if there were no selection maintaining the *pI* of a protein, because of its irrelevance for the proteins function for example, we will expect to observe *pI* shifts of the protein in multiple lineages. Although some proteins show this pattern (Supporting Information Fig. 2, Table III), many other proteins are evolving in a species-specific manner. It is difficult to prove selection on *pI* when the proteins differ widely in sequence length (see Fig. 3), since rate tests of synonymous versus nonsynonymous sites can only be calculated for aligned residues, excluding gaps.

To investigate changes in *pI* that are a consequence of a possible adaptation we applied a classification of amino acids according to their charge used in many studies.[30–34] This classification defines three classes of amino acids, positively charged ones (R, K, H), negatively charged ones (E, D), and conservative ones (all other amino acids). Radical changes are defined as those that change an amino acid from one class to the other, and conservative changes as all other changes. Radical changes will consequently affect the *pI* in a more dramatic way than conservative changes.

It is known that there are more conservative amino acid substitutions than radical substitutions in terms of charge in protein evolution.[30–33] This difference in quantity is usually explained by a higher intensity of purifying selection on radical mutations than on conservative mutations. The opposite on the other hand (more radical changes than conservative ones) argues for a possible structural change as a consequence of adaptation to a functional innovation. The renowned example is the duplicated colobine-specific pancreatic ribonuclease protein (RNASE1B), which underwent functional innovation to adapt to the different *pH* levels of the different digestive systems between this vegetarian leaf eating colobine, and human by shifting its *pI* between both species.[35] Indeed this protein has adapted to $pH = 6$ characteristic of the foregut in of this species, as opposed to its ortholog in human that is adapted to the more basic $pH = 7.4$ characteristic of the small intestine. Because we are only considering the extreme cases of radical over conservative changes, this will reduce the effect of the possible inter-species mutational rate differences.

We applied this method to human and chimp as well as human and mouse pairwise comparisons. On average, the ratio of radical to conservative replacements over the human-mouse orthologs is 0.35. 0.33% (56 proteins) had a value in excess of 1.3. We found 306 orthologous pairs of proteins between human and chimp that show a ratio of radical to conservative changes equal to or higher than 1.3, with a difference in length that is not exceeding 30 amino acids (Supporting Information file 1A). The same method yielded 22 proteins with identical criteria between human and mouse (Supporting Information file 1B). Among the *pI* shifting proteins between human and mouse that are located in the mitochondria, only 3 have experienced length differences higher than 30 amino acids (these are ***ENSG00000055950*** with gaps totaling 79, ***ENSG00000173085*** with a gaps totaling 48, ***ENSG00000129282*** with gaps totaling 34).

To find example of proteins that have shifted their *pI* most likely by evolutionary adaptation we only focused on the proteins that do not show indel differences between orthologs (gaps in Supporting Information file 1 indicates the sizes of indels). Figure 4-A illustrates the example of the Stathmin-4 protein (also known as RB3) that shows 7 substitutions between human and chimp, of which 6 are radical changes. These changes have shifted the protein's *pI* from 6.5 in human to 6 in chimp, and the charge in physiological *pH* from -0.2 to -2.9, respectively. All radical changes seem to be enriching the positive charge in both human and chimp, but human has acquired a higher charge with R and K, as opposed to chimp with only H which is less charged than the other two. All these positions are predicted to occupy exposed solvent accessible residues in the protein (See methods). Moreover, a prediction of the secondary structure of these proteins highlights a shift from predicted helix in human to coil in chimp [Fig. 4(A), See

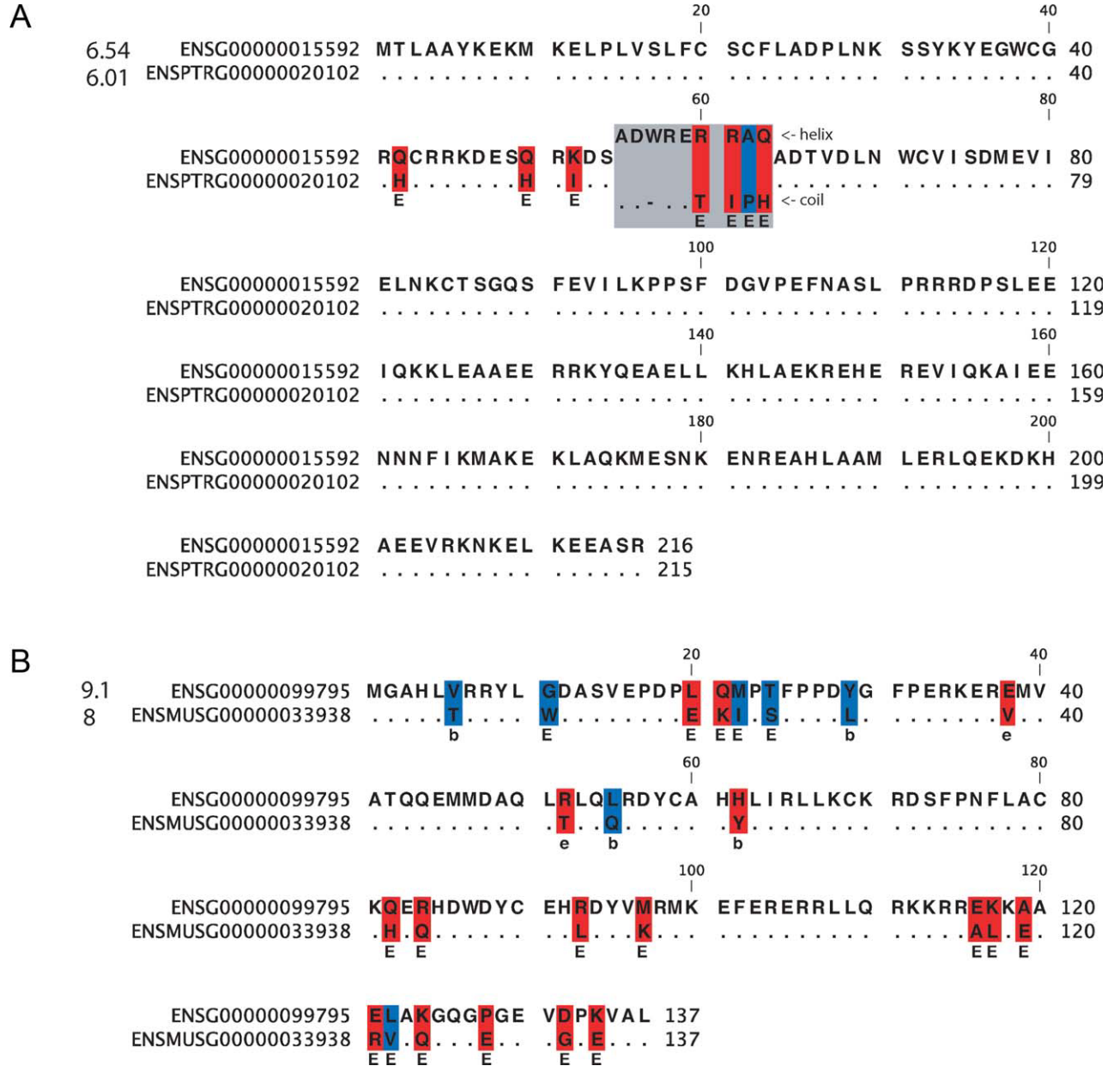

Figure 4

Examples of orthologous proteins presenting high shifts in *pI* accompanied with high radical charge change and low conservative change. The alignment of the proteins was performed with t_coffee. The *pI* values of the proteins (minus their signal peptide) is provided in front of the protein's ID. Dots in the alignment correspond to conservation of the amino acid in both orthologs. Radical changes affecting the charged residues are highlighted in red rectangles. Conservative changes that do not affect the charge of the sequences are in blue rectangles. The letters E, e, and b found under some residues, present respectively Exposed, partially exposed, and buried to solvent accessibility predicted by distill.[36] The grey rectangle highlights differences in secondary structure due to radical and conservative changes. **A**: Alignment of pair of orthologous stathmin-4 proteins between human and chimp. This sequence contains 7 substitutions; of these 6 are radical changes (red) affecting the charge of the protein and only 1 conservative change (blue). All these changes occupy residues that are exposed to solvent accessibility. The prediction of the secondary structure of both sequences showed that the 3 radical changes, and the one conservative change has changed the structure between both sequences, while all the other parts of the protein conserve a similar secondary structure. The structure in this part changed from helix in human to coil in chimp. **B**: Alignment of pair of orthologous NADH dehydrogenase proteins between human and mouse. We have a total of 24 substitutions, of which 17 are radical and 7 are conservative. Sixteen of the radical charges and 4 of the conservative changes are exposed to solvent. The secondary structure is conserved between both orthologs.

methods]. The Stathmin-4 protein (RB3) is a protein that binds to soluble tubulin and thus inhibits the formation and/or increases the depolymerization of microtubules.[37] It is specifically expressed in the cells of the nervous system and acts at the level of neuronal differentiation and activity by regulating the microtubules' dynamics.[37] In our analyses, we used the isoform 2 of the protein that has an additional exon (27 amino acids) between exons 1 and 2 compared with isoform 1. The radical substitutions responsible for the shift in *pI* are located in the region that seems to target and interact with the Golgi membranes.[37]

Is the observed shift in *pI* of RB3, which is mainly affecting charged residues, rare in the human-chimp proteome? Or are there many such instances, in which case it makes it more difficult to argue for selection. We searched for proteins that present similar changes to the RB3 protein (A ratio of radical over conservative changes that is greater than or equal to the value of 1.3 seen for RB3; a difference of radical to conservative replacements of at least 6 or more; a number of gaps that is between 0 and 2, since RB3 has 0 gaps) in the human-chimp set of 18661 orthologs. We only detected 0.1% proteins that presented such a scenario, and only 0.03% that have shifted in their *pI*. This result confirms once more the conservative mode of substitution in mammals.

The shift in *pI* of such an important protein that has been conserved in all mammals might come as an adaptation to changes in functionality or binding partners that may be different between human and chimp. The specific expression of this protein in the cells of the nervous system is of interest in this regard. Given the *pI* in other mammals of this protein, we find that the change in *pI* and charge has most likely occurred in the chimp lineage (the *pI* in monkey of this protein is identical to the human $pI = 6.54$).

Our second example is the NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 7 protein illustrated in Figure 4-B, shows more substitutions between human and mouse, 24 in total, of which 17 are radical. These changes contributed to a net change in charge at physiological *pH* from 5 in human to 2 in mouse, and a shift in *pI* from 9.1 in human to 8 in mouse. Three radical substitutions representing a gain of positive charge in human explains the difference in *pI*. Similarly 16 of the radical changes are predicted to occupy exposed solvent accessible regions (Figure 4-B). However, no likely difference in secondary structure was predicted between the respective orthologs (Figure 4-B). The NADH dehydrogenase [ubiquitine] 1 beta subcomplex subunit 7 is a protein that is part of the complex I of the electron transfer chain in the mitochondria.[38] The diversity of the composition of this complex has been reported among Eukaryotes as well as the gain of additional activities in some species.[38] The shift in *pI* might allow other enzymes to bind to this protein and thus permit the acquisition of new functions for the complex. In cow, for example, an enzyme of complex I has been shown to be also involved in the synthesis of type II fatty acids.[38]

These results put together show that a great deal of *pI* changing proteins experience indels, which consequently alters their *pI*. However, for the group of proteins above, it seems that selection has altered the *pI* by adaptive replacements of certain types of amino acids.

A shift in pI is not significantly accompanied with change in subcellular location

We were curious to know if the shift in *pI* was related to the subcellular localization, and used all the *pI* shifting proteins including those affected by indels, as there were a lack of location information for the smaller set. Our results indicate that the shift in *pI* of orthologous proteins is not more likely accompanied with a shift in location between cellular compartments ($P = 0.33$; proportion test). It was not obvious to automate a comparison on the general GO-terms that were provided so we further categorized the location (see methods). Our method is very stringent to avoid false positives. We found that of the 12830 orthologs between human and mouse for which we know the subcellular localization in both species, 182 proteins (1.4%) were annotated as changing or acquiring a new compartment. For the *pI* shifting proteins between human and mouse (1368 pairwise locations known) we found 24 proteins (1.7%) that have changed location. This result may reflect recent work that has shown that *pI* shows a weak correlation with subcellular localization.[39] However, our method is only considering shifts between compartments and does not take into account the possible cytoplasmic *pI* related stratification of the eukaryotic cell.[8]

pI coevolution between interacting partners

We were interested in knowing if the shift in *pI* of a protein may have resulted from the shift in the *pI* of its interacting partners. We used the BioGRID[40] human datasets to detect sets of interacting proteins that have coevolved in their *pI*. We could only find interaction data for 1696 proteins where the two official symbols provided in BioGRID have a corresponding Ensembl ID. Among this set we detected 396 proteins that have coevolved their *pI* shift with their interacting partner when compared with chimp (Supporting Information file 3). This coevolution does not explain all the *pI* shifts, but is a contributory factor and shows that the coevolution between the *pI*s of interacting proteins is a selective means to shift the *pI* of a protein.

CONCLUSIONS

Isoelectric point of orthologous mammalian proteins are for the most part close between species (see Figure 1). Where there are shifts in *pI*, many of these are because of indels (see Figure 3). Indels are a major player

in generating genetic variation, which if inserted in the coding part will influence the shape and function of the protein. It is difficult to distinguish whether an indel is adaptive or neutral, and whether the *pI* shift associated with the indel is itself adaptive or neutral. This is because we can only state functional divergence if an indel is inserted in regions of functional importance to the protein, which will be interesting to investigate in further studies. However, it is noteworthy to observe that there are more proteins that acquire indels without affecting their *pI* than the opposite (855 out of 1591) which might argue that there is a selection maintaining the *pI* of a protein and that most indels are retained in the protein only because they do not affect its *pI*. However, in some proteins that do not alter their length, specific replacements have altered the *pI*, consistent with selection driving the change in *pI*. The selection could be an adaptation to a new function, new binding partner, or new environment (Figure 4; Supporting Information file 1). But in this work we are unable to identify a universal reason behind the selection on shifting the *pI*. It is entirely possible that selection is driven by multiple factors one of which might be the coevolution with its interaction partner(s), as shown in Supporting Information file 3. Although our analysis focused on global *pI* changes, it is clear that from a perspective of investigating evolution of interacting proteins, further analyses of evolutionary changes in local *pI* will be of great interest. For Stathmin-4 a number of substitution replacements altering the isoelectric point and charge of the protein occur over a very short period of evolution (Figure 4-A). The expression of this protein in the nervous system, might point to differences in neuronal differentiation between human and chimp. However, this remains a speculation and further experimental research into this observation would be required to validate this suggestion. To our knowledge no evidence of adaptive evolution has been previously reported for this protein.

Certain processes and functions are more affected by *pI* shift than others, including noncoding RNA processing, regulation, immune related proteins, and sex related proteins. Some of these processes and functions have been shown to be the major factors behind the great phenotypic differences between mammals,[23] However, the most striking are mitochondrial proteins that seem to be more prone to change their *pI* between species (Table III). It is hard to interpret the reasons behind the shift in *pI* of mitochondrial proteins. Very little is known about possible differences between mammalian mitochondria that might explain these observed *pI* differences. It is unlikely that interspecies differences in mitochondrial *pH* explains the significant differences observed in *pI* shift between mitochondrial orthologous proteins, mainly for three reasons, (1) Frequently the shift in *pI* is interpreted with respect to cellular *pH*; however, previous work has shown that *pI* is not correlated with either *pH* of optimal stability, or with *pH* of optimal activity[41,42] (2) We have shown above that shift in *pI* is not significantly accompanied with change in sub-cellular location, (3) and finally as discussed above we did not detect directionality in the shift experienced by these proteins. Our results show that the shift in *pI* of the mitochondrial proteins does not seem to be driven by indels, given that only 3 orthologous proteins between human and mouse have experienced insertions greater or equal to 30 amino acids. Besides, very few mitochondrial proteins have a ratio greater or equal to 1.3 of radical over conservative substitutions (Supporting Information file 1), indicating that most changes affecting the *pI* are Y and C amino acid substitutions as opposed to the charged residues R, K, H, E, and D. The reason why so many mitochondrial proteins show shifts in *pI* remains unexplained, and it will be of great interest if the underlying cause can be identified.

ACKNOWLEDGMENTS

The authors thank Niall Haslam for discussions. We would also like to thank the reviewers for their helpful comments and suggestions.

REFERENCES

1. Kundrotas PJ, Alexov E. Electrostatic properties of protein-protein complexes. Biophys J 2006;91:1724–1736.
2. Schuurmans Stekhoven FM, Gorissen MH, Flik G. The isoelectric point, a key to understanding a variety of biochemical problems: a minireview. Fish Physiol Biochem 2008;34:1–8.
3. Schwartz R, Ting CS, King J. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res 2001;11:703–709.
4. Garcia-Moreno B. Adaptations of proteins to cellular and subcellular pH. J Biol 2009;8:98.
5. Nandi S, Mehra N, Lynn AM, Bhattacharya A. Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution. BMC Genomics 2005;6:116.
6. Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, Cebrat S. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics 2007;8:163.
7. Seshadri R, Paulsen IT, Eisen JA, Read TD, Nelson KE, Nelson WC, Ward NL, Tettelin H, Davidsen TM, Beanan MJ, Deboy RT, Daugherty SC, Brinkac LM, Madupu R, Dodson RJ, Khouri HM, Lee KH, Carty HA, Scanlan D, Heinzen RA, Thompson HA, Samuel JE, Fraser CM, Heidelberg JF. Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. Proc Natl Acad Sci USA 2003; 100:5455–5460.
8. Flegr J. A possible role of intracellular isoelectric focusing in the evolution of eukaryotic cells and multicellular organisms. J Mol Evol 2009.
9. Bartik K, Redfield C, Dobson CM. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional 1H NMR. Biophys J 1994;66:1180–1184.
10. McIntosh LP, Hand G, Johnson PE, Joshi MD, Korner M, Plesniak LA, Ziser L, Wakarchuk WW, Withers SG. The pKa of the general acid/base carboxyl group of a glycosidase cycles during catalysis: a 13C-NMR study of bacillus circulans xylanase. Biochemistry 1996; 35:9958–9966.

11. Nielsen JE. Analyzing enzymatic pH activity profiles and protein titration curves using structure-based pKa calculations and titration curve fitting. Methods Enzymol 2009;454:233–258.
12. Davies MN, Toseland CP, Moss DS, Flower DR. Benchmarking pK(a) prediction. BMC Biochem 2006;7:18.
13. Nielsen JE, McCammon JA. On the evaluation and optimization of protein X-ray structures for pKa calculations. Protein Sci 2003;12:313–326.
14. Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 1998;6:122–130.
15. Diella F, Gould CM, Chica C, Via A, Gibson TJ. Phospho.ELM: a database of phosphorylation sites–update 2008. Nucleic Acids Res 2008;36:D240–D244.
16. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 2004;4:1633–1649.
17. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 2009;10:48.
18. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLo S Comput Biol 2007;3:e39.
19. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
20. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–1591.
21. Pryor PR, Luzio JP. Delivery of endocytosed membrane proteins to the lysosome. Biochim Biophys Acta 2009;1793:615–624.
22. Rens W, Grutzner F, O'Brien PC, Fairclough H, Graves JA, Ferguson-Smith MA. Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. Proc Natl Acad Sci USA 2004;101:16257–16261.
23. Levine M, Tjian R. Transcription regulation and animal diversity. Nature 2003;424:147–151.
24. Cui Q, Yu Z, Purisima EO, Wang E. MicroRNA regulation and interspecific variation of gene expression. Trends Genet 2007;23:372–375.
25. Reiner R, Ben-Asouli Y, Krilovetzky I, Jarrous N. A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription. Genes Dev 2006;20:1621–1635.
26. Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nat Struct Mol Biol 2004;11:822–829.
27. Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, Ohta K. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 2008;456:130–134.
28. Hartl DL, Jones EW. Genetics: analysis of genes and genomes. Sudbury, Massachusetts: Jones and Bartlett Publishers; 2004. 854 p.
29. Tompa P. Structure and function of intrinsically disordered proteins. Boca Raton, FL: Chapman & Hall/CRC; 2009. 359 p.
30. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. New York: Academic Press; 1965. pp 97–116.
31. Epstein CJ. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 1967;215:355–359.
32. Clarke B. Selective constraints on amino-acid substitutions during the evolution of proteins. Nature 1970;228:159–160.
33. Dayhoff MO, Eck RV, Park CM. A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, MD; 1972. pp 89–100.
34. Zhang J. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 2000;50:56–68.
35. Zhang J, Zhang YP, Rosenberg HF. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 2002;30:411–415.
36. Bau D, Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G. Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 2006;7:402.
37. Curmi PA, Gavet O, Charbaut E, Ozon S, Lachkar-Colmerauer S, Manceau V, Siavoshian S, Maucuer A, Sobel A. Stathmin and its phosphoprotein family: general properties, biochemical and functional interaction with tubulin. Cell Struct Funct 1999;24:345–357.
38. Remacle C, Barbieri MR, Cardol P, Hamel PP. Eukaryotic complex I: functional diversity and experimental systems to unravel the assembly process. Mol Genet Genomics 2008;280:93–110.
39. Chan P, Warwicker J. Evidence for the adaptation of protein pH-dependence to subcellular pH. BMC Biol 2009;7:69.
40. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–D539.
41. Talley K, Alexov E. On the pH-optimum of activity and stability of proteins. Proteins 2010;78:2699–2706.
42. Alexov E. Numerical calculations of the pH of maximal protein stabilty. The effect of the sequence composition and three-dimentional structure. Eur J Biochem 2004;271:173–185.