1. Introduction

Insertions and deletions (indels) attract increasing interest because they prove to be useful in phylogenetic reconstruction and play a major role in genomic evolution. Indels have been credited with relatively low levels of homoplasy compared to nucleotide substitutions (Simmons et al., 2001), but this may be true only for the best studied indels in noncoding sequences, where most indels occur, e.g., in birds the incidence of indels in the coding sequences is only 10% of that in the noncoding DNA (Brandström and Ellegren, 2007). Coding sequence indels that affect protein structure are obviously subject to functional constraints (Wolf and Ellegren, 2007). However, at least some non-coding sequences may also be subject to selective pressures (Bird et al., 2006) and thus their indels may not be immune to homoplasy.

Different kinds of indels have been thought to differ in their sensitivity to homoplasy. First, large indels have been expected to carry more phylogenetic information (Fain and Houde, 2004), although this has earlier been disputed by Simmons et al. (2001). Second, the deletion bias alone should increase the probability of homoplasmous deletions as compared to insertions by the very fact that deletions are more frequent than are insertions, and thus the latter should be given more phylogenetic weight (Edwards et al., 2005). However, only limited empirical support has been provided for the first hypothesis and none for the second.

Major progress has been achieved in resolving the relationships among extant birds (Ericson et al., 2006; Hackett et al., 2008), in particular within the Neavases that are subdivided into a well-supported clade Coronaves and weakly supported clade Metaves. Indels provided strong phylogenetic markers in a wide range of organisms including birds (Fain and Houde, 2004; Edwards et al., 2005; Chojnowski et al., 2008). Indels provided strong support for the monophyly of non-struthioniform paleognaths (Harshman et al., 2008), galloanserines (Fain and Houde, 2004), and Neavases (Groth and Barrowclough, 1999; Fain and Houde, 2004). Indels in β-fibrinogen intron 7 showed a very high consistency index (CI = 0.91) in the phylogeny of Columbidae (Johnson, 2004) and four of them increased bootstrap support for the subdivision of Neavases into Coronaves and Metaves but their segregation between these two proposed clades was not fully consistent, suggesting some homoplasy (Fain and Houde, 2004). Well-established monophyly of eight ordinal-level neavavian clades, including [kagu, sunbittern] and [swifts, hummingbirds], is supported by intron indels in the clathrin heavy chain genes (Chojnowski et al., 2008) and the monophyly of Apodiformes (Johansson et al., 2001) with
Aegothelidae as their basal branch (Fidler et al., 2004; Barrowclough et al., 2006) was confirmed by indels in three genes. Ericson et al. (2003) found in a myoglobin intron one indel synapomorphic for all Lari and one separating the majority of Lari from the Galericulidae. Also, indels helped establish relationships within the uniform orders of Psittaciformes (de Koot and de Koot, 2005) and Passeriformes (Ericson et al., 2000; Spicer and Dunipace, 2004) and even more uniform passeriform clades such as the sylvioid warblers (Alström et al., 2006).

Avian genomes have long been known to be smaller than those of most other vertebrates (Gregory, 2004a) except for the other two groups of flying amniotes, bats and pterosaurs (Organ and Shedlock, 2009; Smith and Gregory, 2009). The size of avian genome is only 30–40% of that of mammals (Brandström and Ellegen, 2007). The reduction of genome size is attributed primarily to the loss of interspersed repetitive elements and probably preceded the origin of flight in the evolution of archosaurs (Organ et al., 2007). Indels (up to 400 bp) have been implicated in the evolution of genome size (Oliver et al., 2007) but deletion bias is deemed too weak to be a primary determinant of genome size variation (Gregory, 2004b) although its impact on the avian genome has yet to be quantified. Whatever their effect on genome size, indels clearly influence gene length, primarily by shrinking or expanding introns. The net impact of indels on gene size ultimately depends on the frequency and the length of insertions and deletions, and thus a deletion bias alone does not necessarily lead to gene contraction.

For example, there is a strong deletion bias of 3.0 in the introns of 1.23 is much greater than unity as the insertions are on the average ca. 3.7 times longer than are deletions (Matthee et al., 2007).

We sequenced three nuclear gene fragments and, using both substitution and indel data, confirmed the relationships within a major subset of the neognathous clades including the majority of landbirds, waterbirds and raptors, as recovered by Ericson et al. (2006) and Hackett et al. (2008). Subsequently, we studied the distribution of indels on a well-supported tree in order to obtain new information on their utility for phylogenetic reconstruction and their significance in genomic evolution.

2. Materials and methods

2.1. Material

We sequenced fragments of three genes: 28S rDNA, ornithine decarboxylase (ODC), and β-fibrinogen (β-FIB) in 42 species representing 29 families of neognathous birds, and one palaeognathous species as an outgroup (Table 1). Whenever possible, we sampled at least two terminal taxa from a family level group to minimize long branch attraction between distant families.

In the 28S gene we sequenced segments 13–15 that constitute its most upstream part, corresponding to nucleotides 3429 through 4200 of the all rDNA array in the map for Mus musculus (Hillis and Dixon, 1991; Palumbi, 1996). For amplifications we used the universal primers, 28v-50 and 28Sjj-3 reverse.

In the β-FIB gene we sequenced the entire (last) intron 7 with minor fragments of exons 7 and 8. For amplifications we used the primers FIB-B7U-forward and FIB-B7L-reverse (Pyrichtiko and Moore, 1997). Our 926-nucleotide pre-alignment sequence of Gallus gallus bankiva corresponds to nucleotides 21,396,564 through 21,397,489 of Gallus gallus chromosome 4 (NCBI Gallus_gallus-2.1 genome map, NCBI Reference Sequence: NC_006090). In G. gallus bankiva, intron 7 comprises 893 nucleotides (96.4% of the obtained sequence) and the flanking exons 7 and 8 are represented only by 33 nucleotides (3.6% of the obtained sequence). The trimmed sequences of β-FIB that we used for analyses represent effectively intron 7 only.

For ODC amplifications we used the primers OD6-forward and OD8-reverse and, in many cases, an additional pair of internal primers OD6int-forward and OD8int-reverse (Allen and Omland, 2003). Our 706-nucleotide pre-alignment sequence of Tragopan temminckii corresponds to nucleotides 99,664,489 through 99,665,196 of G. gallus chromosome 3 (NCBI Gallus_gallus-2.1 genome map, NCBI Reference Sequence: NC_006090.2). In G. gallus, gene ODC comprises 11 exons and 10 introns (NCBI Reference Sequence: XM_419949.2). As mapped using BLAST, our sequence from T. temminckii comprises entire introns 5 and 6 (jointly 461 nucleotides), the ends of exons 5 (111 nucleotides) and 7 (52 nucleotides), and the entire exon 6 (82 nucleotides). Exon sequences make up 34.7% and intron sequences 65.3% of our sequenced fragment.

There seems to be some variation in the structure of the ODC gene, which confuses the numbering of exons and introns. Allen and Omland (2003) sequenced the ODC gene fragment AF491996.1 in Icterus spurius and were first to use it in avian phylogenetics. When mapped on chromosome 3 of the zebra finch Taeniopygia guttata, another passerine that has only 10 exons and 9 introns in the ODC gene (NCBI Reference Sequence: XM_002198078.1), the AF491996.1 sequence covers exons 4 through 6 and complete introns 4 and 5. However, when mapped on chromosome 3 of G. gallus, the AF491996.1 sequence covers exons 5 through 7 and complete introns 5 and 6. Thus, the exon and intron numbers in I. spurius do not correspond to those in either G. gallus (with the phase shift of 1) or T. guttata (with the phase shift of 2).

Aside from the names, most important for the phylogenetic purposes is that all sequenced fragments are orthologous rather than paralogous. There is no evidence for β-FIB gene paralogues in birds (Morgan-Richards et al., 2008) and there is only one copy of true ODC gene. The only avian ODC parologue is ODCp/AZIB that belongs to a distant family (Ivanov et al., 2010). Our simulations using ODCp/AZIB sequence of G. gallus (NCBI G. gallus-2.1 genome map, NCBI Reference Sequence: NW_001471654.1) show, that its sequence differs in about 51% from the true ODC gene and does not bind the primers OD6 and OD8 that we used for amplification. The 28S fragment occurs in multiple copies all of which are subject to concerted evolution (Hillis and Dixon, 1991) that is slow compared to the other two genes and thus the influence of this gene on our final topology is weak. Our matrix is therefore unlikely to include data from any paralogues rather than orthologues and thus we do not expect discrepancies between the phylogenies of species and genes we used. Hence we used as a reference the best species tree (Fig. 1) rather than locus specific phylogenies.

2.2. DNA extraction, amplification and sequencing

All sequences used in this study have been obtained de novo. Total genomic DNA was extracted mostly from skeletal muscles or blood samples, and only in a few cases from feathers, using modifications of phenol–chloroform method (Laird et al., 1991) or Qia-gen DNAeasy Tissue Kit following the manufacturer protocols. The target DNA fragments were amplified in most cases using ABI GeneAmp 9600 thermocycler and Qia-gen Taq PCR Core Kit following manufacturer protocols. All three genes were subjected to (with slight modifications on the case-by-case basis) an initial denaturation (5 min at 95°C); 35 cycles of denaturation (40 s at 95°C), annealing (40 s at 52°C for 28S, 57°C for β-FIB, 59°C for ODC), extension (90 s at 72°C); and a final extension (8 min at 72°C).
The sequences were pre-aligned in Clustal W with interface BioX1.5 and with default settings (gap opening penalty 10, gap extension penalty 5, gap separation distance 8, delay divergent sequences 30%) and without additional settings. The pre-aligned sequences were trimmed and aligned in the SOAP1.1 program for multiple alignments (Löytynoja and Milinkovitch, 2001) using different combinations of gap opening (from 5 to 12 with step 1) and gap extension (from 6 to 9 with step 0.5) penalties. The obtained sequences were trimmed and aligned in the Clustal W program with default settings (gap opening penalty 10, gap extension penalty 5, gap separation distance 8, delay divergent sequences 30%) and without additional settings. The pre-aligned sequences were trimmed and aligned in the SOAP1.1 program for multiple alignments (Löytynoja and Milinkovitch, 2001) using different combinations of gap opening (from 5 to 12 with step 1) and gap extension (from 6 to 9 with step 0.5) penalties. The obtained alignment matrix was used for a strict consensus alignment without unstable nucleotide blocks (ambiguously aligned positions) that leave room for nonhomologous matches. Consensus alignments were viewed and edited in 4SALE (Seibel et al., 2006) and submitted to subsequent analyses as nexus files. Indels were coded as binary characters using SeeState1.22 program (Müller, 2005) and the simple indel coding (sic) algorithm (Simmons and Ochoterena, 2000; Simmons et al., 2007) which is approximately as effective as the simple indel coding (sic) algorithm (Simmons and Ochoterena, 2000; Simmons et al., 2007) which is approximately as effective
as the modified multiple indel coding (Simmons et al., 2007). The sequence fragments with indels were removed prior to the Bayesian analysis of substitution partitions if a gap was present in over 50% of sequences in the alignment.

2.4. Phylogenetic analyses

We used the ostrich (Struthio camelus) as an outgroup. In order to avert long branch attraction, we based our phylogenetic analysis primarily on Bayesian inference (Huelsenbeck et al., 2001, 2002). Six data partitions were defined a priori, two for each of the three sequenced gene fragments – one for substitutions and one for indels. We did not use codon position partitioning because protein coding sequences were effectively absent from two genes (β-FIB and 28S) and represented only about one third of the ODC sequence. In order to determine the contribution of indels to the resolution of phylogeny we performed three Bayesian analyses: for all six partitions jointly (Fig. 1: SUBIND tree), for three substitution partitions alone (Fig. 2: SUB tree), and for three indel partitions alone (Fig. 3: IND tree). In addition, we performed three locus specific analyses for substitutions (Fig. S1: β -FIB tree, Fig. S2: ODC tree, Fig. S3: 28S tree).

The models of nucleotide substitutions for Bayesian analyses were selected individually for each partition using the Bayesian Information Criterion (BIC) and two Akaike Information Criterion (AIC) variants implemented in MultiPhyl program (Keane et al., 2007) that can test 56 currently available models of sequence evolution. The BIC test was considered decisive in cases of conflict. The best models for single gene partitions were HKY + I for 28S, TVM + G for β-FIB, and TVM + I + G for ODC. However, as TVM models cannot be routinely executed by any of the programs that are broadly used for Bayesian analysis, they were replaced, respectively, by GTR + G and GTR + I + G models that have one free parameter extra (in the MrBayes script these models are: 28S – nst = 2 rates = propinv; Fibr – nst = 6 rates = gamma and ODC – nst = 6 rates = invgamma). For indel partitions we applied Bayesian analysis using the model implemented in MrBayes for standard discrete data and based on the Mk model (Lewis, 2001).

Regardless of the model used, all analyses were performed using the same logic and procedures of Bayesian inference as implemented in MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001). To reduce the chance of reaching the apparent stationarity on local optima, two separate runs consisting of four Markov chains for each analysis were performed (in every case three chains were cold and one heated, as a default in MrBayes). Each chain was performed by $20 \times 10^6$ generations and was sampled every 100 generations. The assumptions were congruent with the default settings to a random starting tree, priors constraining the same set of branch lengths for each partition, and heating values of Markov chains. Stationarity and convergence of analyses was estimated

![Fig. 1. Bayesian 50% majority rule tree from three substitution partitions and three indel sic-coding partitions (SUBIND). Branch lengths correspond to number of characters state changes (substitutions and indels).](image-url)
by default MrBayes statistics and graphically in Tracer (Ram‐
baum and Drummond, 2007). Burn-in trees and parameters were dis‐
carded (5000 samples or 0.5 × 10⁶ generations in every case) and
the remaining trees and associated parameters saved, with the fre‐
quency of clades representing estimation of posterior probabilities.
All phylogenetic reconstructions were generated using MrBayes
3.12 parallel version that was run on WCSS supercomputers in
the operational environment of ScientificLinux.

2.5. Indel analyses

In order to determine the significance of indels as both phylo‐
genetic markers and indicators of genomic evolution we compared the
SUB, IND, and SUBIND trees and mapped the indels onto the SUBIND
tree. The MacClade 4.08 parsimony criterion (Maddison and Maddi‐
sion, 1992) was used to reconstruct the number of indel events
(characters), and to determine their states that is, to distinguish be‐
tween insertions or deletions. Characters were assumed as un‐
weighted, unordered, and thus reversible (Fitch parsimony). All
most parsimonious reconstructions of character evolution were
examined for each character. Only unambiguously reconstructed
insertions and deletions (206 out of the total of 287 indels) were
used in further analyses (a change was counted as unambiguous
only when it occurred on the same branch in all most parsimonious
reconstructions of character evolution). We determined the number
of synapomorphic indels supporting each clade (Fig. 4) and calcu‐
lated the levels of homoplasy as indicated by standard fit indices,
consistency index (CI), retention index (RI), and rescaled consistency
index (RC), that were calculated for each indel separately (single‐
character indices) and for all indels in a gene fragment jointly
(ensemble indices). When calculating individual fit indices for pola‐
rized indels, we categorized as deletions only those indels whose all
state changes were deletions, and as insertions only those indels
whose all state changes were insertions. Indels with state changes
including any combination of insertion(s) and deletion(s) have been
excluded from the analysis. Based on the obtained figures, we deter‐
rmined a correlation of the length of parsimony informative indels
with their fit indices (nonparametric Spearman R) and the differ‐
ces in mean fit indices of such indels between the three genes
(Kruskal–Wallis ANOVA or K–W test and median test).

We analyzed the entire set of 287 unpolarized indel characters,
which is immune to potential errors of the accepted topology
(Fig. 3), and, separately insertions and deletions in that topology.
Single insertions and deletions were treated as separate indel char‐
acters when analyzing the set of polarized indels, the number of
which is a sum of unambiguously polarized indels and the number
of extra (more than one) times they occur (as homoplasies) in the
accepted topology. We determined: (1) the numbers and length
frequencies of unpolarized indels in each gene and in all three
genes jointly by testing their agreement with the normal distribu‐
tion (Kolmogorov–Smirnov–Lilliefors or K–S–L test, Shapiro–Wilk
or S–W test); (2) the numbers and length frequencies of insertions
and deletions in each gene and in all three genes jointly (K–S–L and
S–W tests); (3) differences in unpolarized indel length between the
three genes (K–W and median tests); (4) differences in insertion
and deletion length separately between the three genes (K–W
and median tests); (5) differences in length between insertions and
deletions for each gene fragment (K–W and median test). All statisti‐
cal tests and characterization of indels were performed using the
In an attempt to determine whether the sequence length of each gene fragment increased or decreased in avian evolution as a result of indel fixation, we compared the mean length values of insertions and deletions in each gene fragment with the confidence intervals for the experimental insertion/deletion ratios. Also, we calculated the indel fixation rates (per one site = nucleotide) in Neoaves and Galloanseres by averaging the numbers of indels from, respectively, 27 neoavian and 2 galloanserine lineages that we defined as leading to separate families. We counted the indels in Fig. 4 from the split Neoaves/Galloanseres through terminal branches. In order to avoid an accidental bias by several families represented by more than one species, we averaged the number of indels from two or three terminal branches, thus reducing their impact on the neoavian or galloanserine average to that of single-species families.

We employed correlation statistics in order to assess whether the distribution of rate heterogeneity was similar among substitutions and indels. First, we compared the branch lengths estimates from the SUB and IND consensus trees by patristic distances approach (903 compared distances in every matrix). We then calculated patristic distance matrices for the SUB and IND trees and tested the correlations between them by Pearson correlation coefficient (r) and Mantel test (100,000 permutations) with the help of PATRISTIC v1.0 and ZT v1.1. Second, we tested correlation between patristic distance matrices (as in the case of SUB and IND trees) and branch length correlation for indel and substitution branch length sets using parsimony reconstructed branch lengths for indels and substitutions on constrained topology SUBIND (MacClade 4.08). Because of moderate sample size (77 branches in every case) and the lack of normal distributions of branch lengths (Shapiro–Wilk test), we adopted a nonparametric Spearman rank correlations R test.

3. Results and discussion

Altogether we sequenced three gene fragments from 43 avian species in 30 families (Table 1). The final data matrix used for Bayesian analyses (after coding and exclusion of indel sequences) comprises 2184 nucleotides (647 bp of 28S, 857 bp of ODC, and 680 bp of β-FIB) and 287 indels (15 of 28S, 128 of β-FIB, and 144 of ODC).

3.1. Phylogeny

For each of six Bayesian analyses including the three main analyses (SUB, IND and SUBIND) and the three analyses for substitutions only in each gene fragment separately, we obtained a sample of 195,000 trees from each run, thus a total of 390,000 sampled trees from one analysis. One SUB, one IND, and one SUBIND Bayesian 50% majority rule consensus tree was obtained from each sample. The estimated marginal likelihoods (lnL) in SUB analysis for tree sample are 18760.37 (arithmetic mean) and 18808.78 (harmonic mean) for run 1 and 18758.85–18822.27 for run 2. The corresponding values for IND analysis are 1520.42–1591.02 (run1) and 1521.85–1588.51 (run2) and for SUBIND analysis 20321.017–20371.66 (run1) and 20320.50–20373.39 (run2).

Fig. 3. Bayesian 50% majority rule tree from three indel sic-coding partitions (IND). Branch lengths correspond to number of indels. Note that the tree for Neoaves differs substantially from the best (SUBIND) tree in Fig. 1 and so does the distribution of indels from that in Fig. 4.
Fig. 4. All indels mapped on SURIND tree using MacClade 4.08 parsimony. Branch lengths proportional to number of indels, hence the nodes not separated by at least one indel have been collapsed, resulting in polytomies as well as accidental changes in the sequential order of taxa. Two ambiguous changes at the root of Coronaves are from β-FIB and ODC, and the one at the root of Metaves is from β-FIB gene.
The estimated marginal likelihoods (InL) for β-FIB analysis are: −10125.12 (arithmetic mean)−10176.97 (harmonic mean) for run 1 and −10125.43−10173.16 for run 2. The values for ODC analysis are: −7094.53−7142.65 (run1) and −7094.59−7150.15 (run2) and for 28S analysis: −1300.87−1386.33 (run1) and −1304.92−1392.19 (run2). The convergence diagnostic (PSRF or potential scale reduction factor) of 1.00 for all free parameters of the model, 33 for SUB, 4 for IND, 36 for SUBINF, 12 for β-FIB, 13 for ODC, and 7 for 28S indicates a good sample from the posterior probability distribution.

The best supported, SUBIND tree (Fig. 1) that is based on both substitutions and indels, is in general agreement with the topologies obtained by both Ericson et al. (2006) and Hackett et al. (2008). While the tree recovered from indels alone (Fig. 3) is less structured than the tree based on substitutions (Fig. 2) neither the substitutions (Fig. 2) nor the indels alone (Fig. 3) yield the subdivision of Neoaves into Metaves and Coronaves. This suggests that the combined information from substitutions and indels generates an added value that allowed us to obtain essentially the same division of Neoaves into Metaves and Coronaves. This suggests that a single indel carries more character weight than a single substitution, although this difference may possibly result from our omission of the cuculiforms and musophagids that may attract the gruiform branch.

Within the Charadriiformes, our topology (Fig. 1) identifies clades [Burhinidae, Laridae] and [Charadriidae, Scolopacidae] with the maximum support (Pp = 1) for both clades (Fig. 3) and thus agrees with the trees obtained by Paton et al. (2003) using RAG-1 gene sequence and Livezey (2010) using phenotypic characters. However, in 12 out of 19 charadriiform published topologies (Livezey, 2010; Mayr, 2011b), the Burhinidae are placed closer to the Charadriidae than to Lari. Fain and Houde (2007) left a polymy in the position of Burhinini unresolved, Chu (1995) placed Burhinidae equidistant to Charadriid and Lari, and one of three Mayr’s (2011b: Fig. 3A) topologies is a polytomy of four clades each comprising one of the four discussed taxa. Since the relationships within the Charadriiformes have yet to be settled, our topology may prove correct despite the paucity of sampled families.

Nearly all the assembled fit indices CI, RI i RC for indels are well over 0.5 and thus much higher than those for substitutions (Table 3). Also Fain and Houde (2004) and Johnson (2004) found in FGB-int7 the consistency indices for indels to be much higher than for the nucleotides. This is also true for all mean indices that are significantly higher for indels (Table 4) than those for substitutions except for 28S indels (the exception being probably due to a small sample size). We suggest that a single indel carries more character weight than a single substitution, although clearly much more work is needed to determine realistic and safe multiplication factors.

No correlation of the fit indices with indel length has been found either as calculated for all indels jointly or for particular gene fragments separately (Table 5). This provides evidence against phylogenetic weighting of indel length, in agreement with Simmons et al. (2001). Fain and Houde (2004; Fig. 2) plotted β-FIB parsimony informative indel length and consistency indices as evidence for a greater weight of longer indels but the correlation of 0.2 (Spearman R) as calculated from their data is rather low although significant (0.05). Analogous correlation calculated from our data for β-FIB parsimony informative indels is even lower (Spearman R = 0.10) and insignificant. This suggests a weak correlation between β-FIB indel length and CI that is insignificant with
our sample of 40 indels but significant with the larger sample of 161 indels in Fain and Houde’s study.

The mean values of all individual fit indices CI, RI and RC are higher for deletions than for insertions for all three genes jointly but the difference is barely significant (K–W test, $p < 0.1$) (Table 4) although it is significant for the ODC fragment alone. Thus our results do not support the hypothesis that deletions are more likely to be homoplasious because of being more frequent than insertions and show that the opposite may be true for some genes.

The phylogenetic utility of indels, as measured by their individual fit indices, does not significantly vary between individual genes (Tables 6 and 7). In contrast, the averaged individual consistency index is significantly higher for substitutions in β-FIB as compared to the ODC fragment (K–W test with post hoc multiple comparisons and median test) although the difference is small (Tables 6 and 7).

### 3.2. Genomics

There are no differences in indel length between the three genes, either for all indels jointly or insertions and deletions separately (Table S1, K–W and median tests). Aside from the only two deletions in 28S, a normal distribution of indel lengths can be rejected at the significance levels of 0.01 (Fig. S1 and Table S3). In fact, the distributions of both insertions and deletions for each fragment reveal a strong right skewness (Table 2, Fig. S1), suggesting an exponential or Zipfian (Yule’s) distribution. The most frequent are short, one- to three-nucleotide indels that contribute over 50% of all indels. Second in frequency are 4–11-nucleotide indels. The least frequent are medium-size indels of 12–20 nucleotides, whereas indels over 20 nucleotides are somewhat more frequent, which results in slightly bimodal distributions (Fig. S1 and Table S3). Such a distribution is consistent with the proposal that different mechanisms are responsible for the generation of short and long indels (Creer, 2007).

The average insertion and deletion lengths in the neognathous birds are, respectively, 3.63 and 6.69 nucleotides (Table 2), the deletion bias is ca 2.5 (147/59), and thus the overall NITD ratio amounts to 0.22, suggesting a rapid decrease in gene size. However, this may not be representative for the entire avian (or at least neognathous) genome, as the deletion bias as calculated by comparing the entire genomes of two phasianids (Gallus and Meleagris) is much lower (1.42) and limited to macrochromosomes (Brandström and Ellegren, 2007).

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembled fit indices (CI_inf – consistency index, RI_inf – retention index, and RC_inf – rescaled consistency index) for parsimony informative indels (left figure) and substitutions (right figure) as mapped on the SUBIND tree (Fig. 1).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CI_inf</th>
<th>RI_inf</th>
<th>RC_inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>28S</td>
<td>0.60/0.38</td>
<td>0.78/0.52</td>
<td>0.47/0.20</td>
</tr>
<tr>
<td>β-FIB</td>
<td>0.73/0.55</td>
<td>0.85/0.53</td>
<td>0.62/0.29</td>
</tr>
<tr>
<td>ODC</td>
<td>0.66/0.49</td>
<td>0.69/0.47</td>
<td>0.46/0.23</td>
</tr>
<tr>
<td>All fragments</td>
<td>0.68/0.52</td>
<td>0.77/0.50</td>
<td>0.53/0.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences between insertions and deletions (left figure) and between all indels and substitutions (right figure) in mean fit indices (CI_inf – consistency index, RI_inf – retention index, and RC_inf – rescaled consistency index, all calculated for parsimony informative characters) for every gene fragment.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CI_inf</th>
<th>RI_inf</th>
<th>RC_inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>28S</td>
<td>-/p = 0.31</td>
<td>-/p = 0.30</td>
<td>-/p = 0.25</td>
</tr>
<tr>
<td>β-FIB</td>
<td>p = 0.52/p = 0.00</td>
<td>p = 0.56/p = 0.00</td>
<td>p = 0.56/p = 0.00</td>
</tr>
<tr>
<td>ODC</td>
<td>p = 0.01/p = 0.00</td>
<td>p = 0.01/p = 0.00</td>
<td>p = 0.01/p = 0.00</td>
</tr>
<tr>
<td>All fragments</td>
<td>p = 0.10/p = 0.00</td>
<td>p = 0.11/p = 0.00</td>
<td>p = 0.12/p = 0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations (Spearman R, all insignificant at 0.05) between the length of parsimony informative indels and their phylogenetic utility as measured by CI_inf – consistency index, RI_inf – retention index, and RC_inf – rescaled consistency index.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CI_inf</th>
<th>RI_inf</th>
<th>RC_inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>All indels</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ODC</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.08</td>
</tr>
<tr>
<td>β-FIB</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>28S</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences in mean fit indices (CI_inf – consistency index, RI_inf – retention index, and RC_inf – rescaled consistency index, all calculated for parsimony informative characters) between indels (left figure) and substitutions (right figure) from three different gene fragments.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CI_inf</th>
<th>RI_inf</th>
<th>RC_inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>K–W test</td>
<td>p = 0.52/p = 0.02</td>
<td>p = 0.5/p = 0.25</td>
<td>p = 0.53/p = 0.32</td>
</tr>
<tr>
<td>Median test</td>
<td>p = 1.00/p = 0.04</td>
<td>p = 1.00/p = 0.57</td>
<td>p = 1.00/p = 0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic statistical description of individual fit indices for parsimony informative polarized indels and substitutions.</td>
</tr>
</tbody>
</table>

For Deletions

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
<th>Std. dev.</th>
<th>Valid N</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.88</td>
<td>0.25</td>
<td>1</td>
<td>0.23</td>
<td>28</td>
</tr>
<tr>
<td>CI_inf</td>
<td>0.78</td>
<td>0.25</td>
<td>1</td>
<td>0.28</td>
<td>807</td>
</tr>
<tr>
<td>RI_inf</td>
<td>0.69</td>
<td>0.25</td>
<td>1</td>
<td>0.41</td>
<td>807</td>
</tr>
<tr>
<td>RC_inf</td>
<td>0.64</td>
<td>0.25</td>
<td>1</td>
<td>0.46</td>
<td>807</td>
</tr>
</tbody>
</table>

For Insertions

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
<th>Std. dev.</th>
<th>Valid N</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.88</td>
<td>0.25</td>
<td>1</td>
<td>0.23</td>
<td>8</td>
</tr>
<tr>
<td>CI_inf</td>
<td>0.89</td>
<td>0.5</td>
<td>1</td>
<td>0.21</td>
<td>481</td>
</tr>
<tr>
<td>RI_inf</td>
<td>0.82</td>
<td>0.5</td>
<td>1</td>
<td>0.37</td>
<td>481</td>
</tr>
<tr>
<td>RC_inf</td>
<td>0.80</td>
<td>0.5</td>
<td>1</td>
<td>0.39</td>
<td>481</td>
</tr>
</tbody>
</table>

For Substitutions

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
<th>Std. dev.</th>
<th>Valid N</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.86</td>
<td>0.25</td>
<td>1</td>
<td>0.23</td>
<td>8</td>
</tr>
<tr>
<td>CI_inf</td>
<td>0.58</td>
<td>0.33</td>
<td>1</td>
<td>0.27</td>
<td>307</td>
</tr>
<tr>
<td>RI_inf</td>
<td>0.40</td>
<td>0.33</td>
<td>1</td>
<td>0.42</td>
<td>307</td>
</tr>
<tr>
<td>RC_inf</td>
<td>0.31</td>
<td>0.33</td>
<td>1</td>
<td>0.43</td>
<td>307</td>
</tr>
</tbody>
</table>
There are no detectable length differences between insertions and deletions in 28S and ODC genes, but such differences are present in the β-FIB intron where deletions are on the average 2.5 times longer, and have the median value 1.5 times higher than insertions (Tables 2 and S2). Accordingly, the distribution of deletions in β-FIB is much more right skewed than that of insertions (Table 2, Fig. S1). In conjunction with the deletion bias of 2.38 (Table 8), the differences in length result in the NITD ratio as low as 0.17 which must lead to a rapid shortening of this intron. Deletion bias has been recorded in β-fibrinogen intron 7 in many avian lineages (Prychitko and Moore, 2003) and turned out to be as high as ca. 6 in the Columbidae with a record low NITD ratio of 0.05 (Johnson, 2004) although this may be heavily exaggerated by microsatellite length variation (Brandström and Ellegren, 2007: p. 1699).

In the ODC fragment, deletions are 3.83 times more frequent than insertions and the difference is statistically significant (Table 8). Without significant differences in length between insertions and deletions (Tables 2 and S2) this must lead to a shortening of this gene with the NITD ratio of 0.21.

In contrast to the other two genes, in the 28S fragment insertions are significantly more frequent (Table 8) and prevail over deletions with the overall NITD ratio of 1.38. This is a slowly evolving gene expanded primarily in the early neovavian evolution with rare deletions occurring only in terminal branches (Fig. 4).

The crown-group neognaths (van Tuinen, 2009) and placentals (therians) (Murphy and Eizirik, 2009) split at about the same time, some 105 mya, and yet the avian (neognathous) indel fixation rate of 0.01385 per site is nearly 2.5 times higher than 0.00559 in placental mammals as determined by Matthee et al. (2007). This difference may possibly be somewhat exaggerated but seems ways too big to be accounted for by differences in methods. Another caveat stems from the fact that mammalian and avian fixation rates were calculated from different nonoverlapping sets of three genes that may differ in indel fixation rates. However, our results provide strong evidence that indel fixation rates are likely to be on the average much higher in neognathous birds than in therian mammals.

We discovered a considerable taxonomic variation in indel fixation rates (Fig. 4). The neovanian rate of 0.01132 indels/site, which is based on a large sample and thus reliable, is nearly 1.5 times lower than the galloanserine rate of 0.01637 indels/site, which is based on a small sample of three species from two modern dominant families (Anatidae and Phasianidae) and thus may be exaggerated. However, even after removing the value for Tragopan as a possible outlier, the rate of 0.01465 indels/site, which is identical with either Gallus or Anas, remains 1.3 times higher than the neovarian average that includes an obvious outlier for the Railidae.

Among the Neoaves, the rates within most family-level and other infraordinal clades (Accipitridae, Cathartidae, Sulaoida, Ciconiidae, Scopidae, and Charadrii) are fairly uniform, but extremely uneven in the Ardeidae, Threskiornithidae, and Gruidae. The most striking are the accelerated rates in the Railidae and Tragopan. Assuming the split between the rails and cranes at 64.5 mya (Houde, 2009) and that between the neovanians and galloanserines at 105 mya (van Tuinen, 2009), the indel fixation rate of 0.0002626 indel/site/my in the Railidae has been over 2.4 times higher than the average rate of 0.0001078 indel/site/my for Neoaves (including the Railidae) and 8.2 times higher than the average for the two Gruidae (Fig. 4). The indel fixation rate of in Tragopan lineage within the Phasianinae (pheasants) has been 3.5 times that in the Gallus (junglefowl, chicken) lineage. Since the junglefowls are only distantly related to the pheasants (Crowe et al., 2006), there is a possibility that other pheasants may share the fast rate with Tragopan. Highly differentiated indel fixation rates have been reported among the placentals, with rodents and afrotherians accumulating indels more than twice as fast as other lineages (Matthee et al., 2007).

There is no obvious explanation for the striking differences of indel fixation rates in either mammals or birds. However, there is good evidence that indel fixation rates correlate with lineage-specific evolutionary rates as determined by other measures. It is known that indels, including those affecting non-coding sequences, may be subject to selective pressures (Bird et al., 2006) and that indels increase mutation rates in the surrounding sequences (Tian et al., 2008). Accordingly, a high indel fixation rate in rodents correlates with elevated rates of change in both nuclear and mitochondrial genes (Matthee et al., 2007).

Our results support a strong association between indel fixation rates and lineage-specific substitution rates as demonstrated by high and significant correlations between SUB tree and IND tree patristic distance matrices (Pearson $r = 0.91$, $p < 0.0001$), between patristic distance matrices for indel and substitution branch length reconstructed on constrained topology SUBIND (Pearson $r = 0.88$, $p < 0.0001$), and for indel and substitution branch length reconstructed on constrained topology SUBIND (Spearman $R = 0.85$, $p < 0.05$).

Highest substitution rates in both Railidae and Phasianidae have been recorded by Hackett et al. (2008) and ourselves (Fig. 2). In addition, the rails are known for their high adaptability to insular habitats and a remarkably fast evolution of flightlessness (Slikas et al., 2002) and the phasianid galliforms show a much higher recombination rates than the passerines (Ellegren, 2007; Stapley et al., 2008). Also, on a smaller taxonomic scale, the striking difference in the number of indels between the bitterns (Botaurus stellaris) and two heron species (Fig. 4) correlates with about two times faster substitution rate in the same genes (Fig. 2) and 1.25 times faster evolution of single-copy nuclear DNA (Sheldon et al., 2000) in the bitterns compared to herons. We suggest that the variation of indel fixation rates may be accounted for in terms of lineage-specific evolutionary rates rather than general, phylogeny-independent allometric relationships. However, while all studied lineages with above-average indel fixation rates also show heightened substitution rates, the opposite is not true as shown by heightened substitution rates in Scopus, Scolopax, the falcons, Eurypyg, and Podiceps (Figs. 1 and 2), that do not reveal any above-average indel fixation rates (Fig. 4).

Acknowledgments

We thank our technicians, K. Nowak and A. Dąbrowska, for sequencing work, and K. Lipiec-Sidor for managing tissue samples which were kindly provided by J. Dean (National Museum of Natural History, Smithsonian Institution), A. Kruszewicz (The Warsaw Zoo), P. Ćwietniak (The Poznań Zoo), and T. Grabiński (The Wrocław Zoo). The manuscript greatly benefitted from comments by three
anonymous reviewers as well as M. Wolsan (Museum and Institute of Zoology, Polish Academy of Sciences). The laboratory part of this project was supported by Poland’s Ministry of Science and Higher Education grant 3 P04 C 06925 and the computational part by Wroclaw Network-Supercomputer Center (WCCS) grant 106/2008.

Appendix A. Supplementary material


References


