Syst. Biol. 62(1):121–133, 2013

DOI:10.1093/sysbio/sys077
Advance Access publication on September 6, 2012

Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study

OMAR ROTA-STABELLI[1,2,*], NICOLAS LARTILLOT[3], HERVÉ PHILIPPE[3], AND DAVIDE PISANI[1,4,*]

[1]*Department of Biology, The National University of Ireland, Maynooth, Co. Kildare, Ireland;* [2]*Sustainable Agro-ecosystems and Bioresources Department, IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy;* [3]*Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montréal, Québec, Canada H3C3J7; and* [4]*School of Earth Sciences and School of Biological Sciences, The University of Bristol, Woodland Road, BS8 1UG Bristol, UK*

**Correspondence to be sent to: School of Earth Sciences and School of Biological Sciences, The University of Bristol, Woodland Road, BS8 1UG Bristol, UK; E-mail: Davide.Pisani@bristol.ac.uk (D.P.); Sustainable Agro-ecosystems and Bioresources Department, IASMA Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige (TN), Italy; E-mail: omar.rota@iasma.it (O.R.-S.).*

Received 12 September 2011; reviews returned 19 December 2011; accepted 30 August 2012
Associate Editor: Peter Forest

Abstract.—Phylogenomic analyses of ancient relationships are usually performed using amino acid data, but it is unclear whether amino acids or nucleotides should be preferred. With the 2-fold aim of addressing this problem and clarifying pancrustacean relationships, we explored the signals in the 62 protein-coding genes carefully assembled by Regier et al. in 2010. With reference to the pancrustaceans, this data set infers a highly supported nucleotide tree that is substantially different to the corresponding, but poorly supported, amino acid one. We show that the discrepancy between the nucleotide-based and the amino acids-based trees is caused by substitutions within synonymous codon families (especially those of serine—TCN and AGY). We show that different arthropod lineages are differentially biased in their usage of serine, arginine, and leucine synonymous codons, and that the serine bias is correlated with the topology derived from the nucleotides, but not the amino acids. We suggest that a parallel, partially compositionally driven, synonymous codon-usage bias affects the nucleotide topology. As substitutions between serine codon families can proceed through threonine or cysteine intermediates, amino acid data sets might also be affected by the serine codon-usage bias. We suggest that a Dayhoff recoding strategy would partially ameliorate the effects of such bias. Although amino acids provide an alternative hypothesis of pancrustacean relationships, neither the nucleotides nor the amino acids version of this data set seems to bring enough genuine phylogenetic information to robustly resolve the relationships within group, which should still be considered unresolved. [Codon-usage bias; nucleotide composition bias; Pancrustacea; phylogenomics; serine; 21-states CAT model; 23-states CAT model.]

NUCLEOTIDES VERSUS AMINO ACIDS

Data-type choice in molecular phylogenetics is of primary importance when dealing with protein-coding genes that can be represented and analyzed at the amino acid (aa), the nucleotide (nt), or the codon level. Amino acid data sets have 20 character states, whereas nucleotide data sets use characters with either 4 (nucleotide) or 61 (codon) states. The main difference between the two data types is that nucleotides accumulate more substitutions than amino acids in the form of synonymous substitutions.

For phylogenetic studies addressing recent divergences, data-type choice is straightforward: nucleotides are more informative than amino acids, because substitutions are more likely to have occurred at synonymous sites (most third-codon positions plus the first positions of synonymous leucine and arginine codons, and first and second positions of synonymous serine codons). For deep-time phylogenetics, the choice is less trivial. Nucleotides still bring more substitutions (synonymous as well as non-synonymous), but because synonymous sites are under minimal (if any) selective constraints, they tend to accumulate substitutions at high rates, leading to saturation over long periods of time. Models of nucleotide evolution can discriminate substitutions representing a reliable source of signal (homologies) from saturation (homoplasies) only to a certain extent. This may either result in poorly resolved phylogenies, or more worryingly, artifactual clades. In addition, because of their relaxed selective constraints, synonymous sites can be under strong mutational pressure, and compositionally driven mutational pressure has previously been shown to be responsible for the biased accumulation of specific nucleotides in unrelated lineages (Foster et al. 1997; Saccone et al. 1999). In such circumstances, the time-homogeneity assumption of stationary models is violated (Lockhart et al. 1992; Galtier and Gouy 1995; Yang and Roberts 1995; Foster 2004; Jermiin et al. 2004; Gibson et al. 2005; Blanquart and Lartillot 2008; Foster et al. 2009), and unless time-heterogeneous models are employed, compositional attractions [see Jeffroy et al. (2006) for review] can sway the results of phylogenetic analyses. Common practice to overcome these problems is to remove third-codon positions from nucleotide data sets, or to use character recoding strategies (e.g., R-Y coding as in Woese et al. 1991). Because saturation and mutational pressures preferentially affect synonymous sites, amino acid data sets are expected to be less prone (Jeffroy et al. 2006; Rota-Stabelli et al. 2010), but not immune (Foster and Hickey 1999), to both problems.

Codon-usage biases can also cause phylogenetic errors in nucleotide-based data sets. (Inagaki and Roger, 2006) and Inagaki et al. (2004) showed that, in the case of deep divergences among the eukaryotic lineages, phylogenetic analyses based on nucleotide sequences of plastid-encoded genes could be misled by unrelated lineages having similar codon-usage biases for

leucine, serine, and arginine. Because codons for leucine, arginine, and serine differ not only at the third position, but also at the first position (leucine and arginine) or at both the first and second positions (serine), codon-usage biases will tend to affect nucleotide data sets, even if the third-codon positions are removed. This bias would naturally extend to 61-state (codon) models (Inagaki and Roger 2006).

Saturation, parallel compositional biases, and parallel codon-usage biases represent 3 problems that might preferentially affect nucleotide data sets, even though some studies (Holder et al. 2008; Seo and Kishino 2009) suggested that, in phylogenetics, nucleotides are preferable to amino acids. However, these conclusions are partly based on simulations that did not account for variations of global nucleotide composition among lineages, and may therefore not directly apply to cases where strong compositional heterogeneities exist. In addition, for deep phylogenetic problems where large sequence alignments are analyzed using Bayesian mixture models (CAT models), amino acids appear to be equally or more accurate than nucleotides (Holder et al. 2008). Furthermore, the above-mentioned studies were based on single-gene analyses and may not apply to phylogenomics.

The Pancrustacea as a Case Study

Uncertainties remain about the internal relationships within the arthropod subphyla, particularly Pancrustacea (Crustacea plus Hexapoda). Crustacea are most likely paraphyletic (Cook et al. 2005; Regier et al. 2005), but the details of their relationships are highly debated. On the basis of respiratory proteins, Ertas et al. (2009) suggested that Remipedia—a small group of anchialine cave crustaceans—are the sister group of Hexapoda. This hypothesis finds support in the neuranatomical analysis of Fanenbruck et al. (2004) and was partially confirmed by Regier et al. (2008) using a matrix of 62 carefully selected protein-coding genes and 12 taxa. Regier et al. (2010), hereafter R2010, dramatically enlarged this taxon sampling to 80 arthropod species. They inferred trees based on both nucleotide sequences (using also a codon model) and the corresponding (i.e., translated) amino acid sequences, finding that many of the crustacean relationships supported by the nucleotides (see also Regier and Zwick 2011) are not found when using amino acids (Regier et al. 2010). Hence, R2010's data set represents an extremely interesting case to evaluate the issue of data-type choice in phylogenomics.

The R2010 nt analyses (under both codon and nucleotide models) identified 6 strongly supported, novel Pancrustacean groups that do not appear in their aa tree (see also Regier and Zwick 2011). More precisely in their nt analyses (Fig. 1a), a monophyletic Remipedia plus Cephalocarida (a group named Xenocarida) is found to be the sister group of Hexapoda in the Miracrustacea clade. Thecostraca plus Malacostraca (Communostraca) and Copepoda form the Multicrustacea, and Branchiopoda is the sister group of Multicrustacea in a clade named Vericrustacea, which, in turn, is the sister group of Miracrustacea. Oligostraca is the sister group of Vericrustacea plus Miracrustacea (a group referred to as Altocrustacea). R2010 took important precautions to avoid systematic errors when performing their nt analyses: (i) they carefully avoided the inclusion of paralogs in their data set, (ii) they used a rich and representative taxon set, (iii) they used a data matrix with only 18% missing data, (iv) they excluded third-codon positions, (v) they excluded the leucine and arginine synonymous codons, and (vi) they tested the use of a codon model. Accordingly, their results (Fig. 1a) should be robust. However, many clades identified in their nt analyses are not supported by their aa analyses (Fig. 1b), which is unexpected and intriguing. In addition, some of R2010's new clades are in discordance with trees derived using both rRNAs and ESTs, which support Branchiopoda, and in some cases Copepoda, to be more closely related to Hexapoda than to Malacostraca (Mallatt and Giribet 2006; Von Reumont et al. 2009; Meusemann et al. 2010; Campbell et al. 2011; Rota-Stabelli et al. 2011).

Here, we addressed the 2-fold problem of data-type choice and crustacean relationships, further exploring the signals in the R2010 data set. We performed a variety of analyses on nt and aa data sets: we used mixture models, we removed all synonymous codons from the nt data set, we used synonymous codons in aa phylogenetics, we performed targeted taxon- and site-sub-sampling experiments, and we performed analyses to identify both compositional and codon-usage biases. We show that different arthropod lineages are differently biased in their usage of synonymous codons, and that these biases could be responsible for the differences observed between the nt and aa analyses.

Materials and Methods

Data Sets

The nucleotide and amino acid data sets.—R2010 data set consists of concatenated regions from 62 nuclear-coded genes across 80 taxa, of which 75 are arthropods. From this data set, R2010 derived several nt alignments implementing different data set treatments. We analyzed the nt data set named noLRall1 + nt2 (80 taxa, 21 823 positions), which excludes all third-codon positions and those first-codon positions encoding one or more leucine or arginine codons. We chose this data set as it is the most adequate (surely the most conservative) to test the effect of data-type choice, as it is the most compositionally homogeneous among those presented by R2010. Furthermore, it provides the highest support to R2010's new pancrustacean groups (Fig. 1a). For our amino acid analyses, we used an aa data set (80 taxa and 13 087 positions) generated by translating the R2010 codon data set. Data sets used in this study have been deposited in Treebase, submission ID 13171. All data sets as well as Supplementary Information can be found

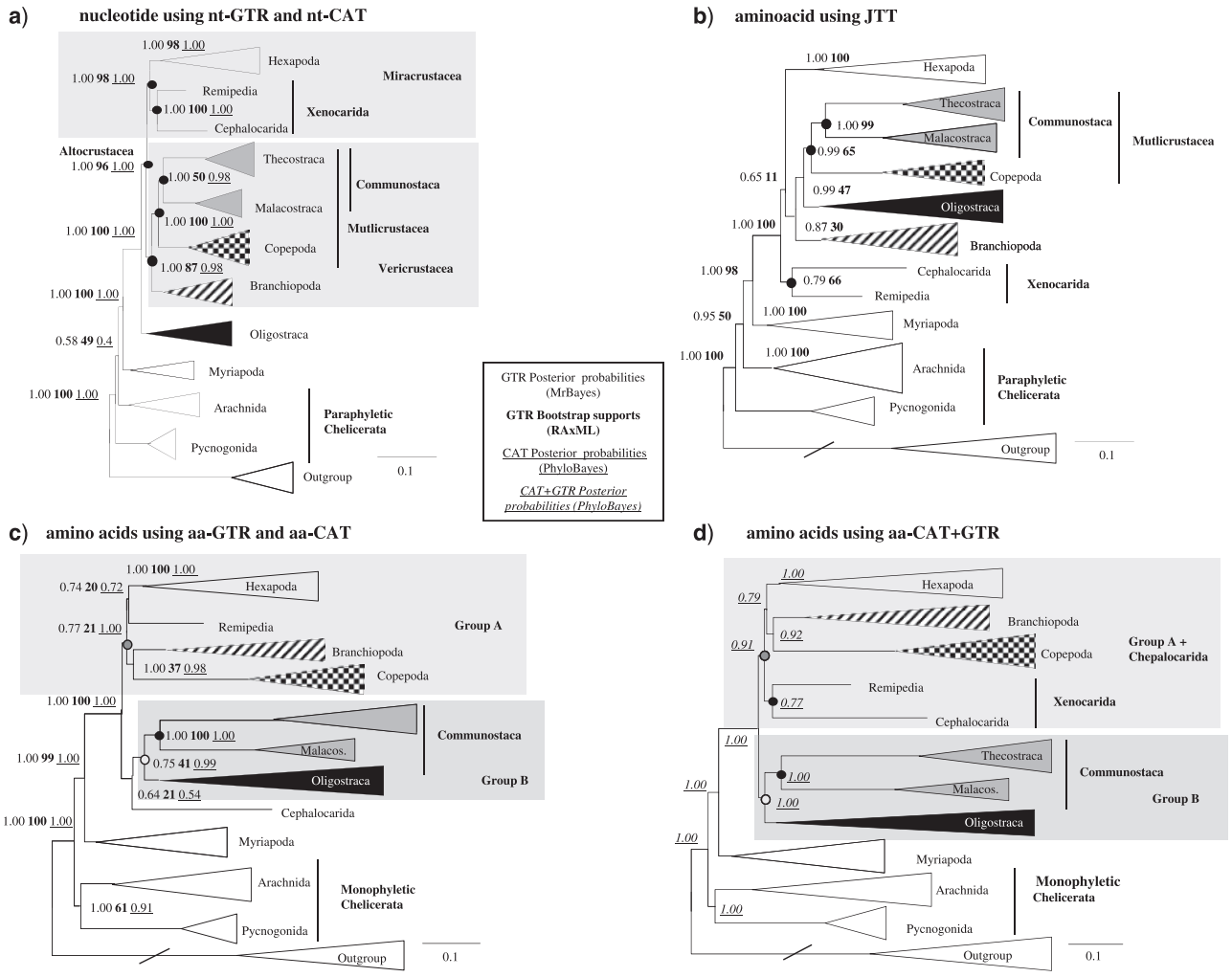

FIGURE 1. Conflict between nucleotides and amino acid data types: a) Analysis of the noLRall1 + nt2 data set under the nt-GTR and the nt-CAT models robustly supports 6 new groups: Vericrustacea, Multicrustacea, Communostraca, Miracrustacea, Altocrustacea, and Xenocarida (black circles at nodes defining the groups). b) Analysis of the corresponding aa data set under JTT does not support 3 of the pancrustacean groups in Figure 1a (Altocrustacea, Miracrustacea, and Vericrustacea). c) Bayesian and ML analyses of the aa data set analyzed using the better fitting aa-GTR and aa-CAT models. These analyses support an alternative view of pancrustacean relationships where Branchiopoda, Copepoda, Hexapoda, and Remipedia are included in a clade named Group A. These analyses strongly support the monophyly of Chelicerata, which was not supported using JTT. d) Bayesian analyses using the best-fitting aa-CATGTR model support a similar tree where Cephalocarida is also a member of Group A. Support values at nodes are posterior probabilities from MrBayes or PhyloBayes, and bootstrap support from Raxml (see legend in the figure and text for details). Lineages have been collapsed for clarity with the length of triangles equal to the longest terminal branch in the collapsed lineage and stem branches equal to the originals. Original trees with full support values are available upon request. The full aa-CAT tree is in Supplementary Figure S1.

in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.7p1k8304.

Tree search parameters applicable to all data sets.—We performed various Bayesian and maximum-likelihood (ML) analyses on both the nt and aa data sets. Parameters that were identically set in all analyses are reported in this section. All ML analyses were performed with RaxmlVI-HPC (Stamatakis 2006), using the fast ML method to obtain bootstrap support (100 replicates). Bayesian analyses were conducted using the MPI version of MrBayes3 (Ronquist and Huelsenbeck 2003; Altekar et al. 2004) and PhyloBayes3 (Lartillot et al. 2009). For both programs, two independent runs were performed and, in the case of MrBayes, each run used 4 differentially heated chains. All Bayesian GTR (Lanave et al. 1984; Yang et al. 1998) and JTT (Jones et al. 1992) analyses were performed using MrBayes. All analyses performed using the CAT and the CATGTR model (Lartillot and Philippe 2004) were performed using Phylobayes. Analyses were stopped when the standard deviation of split frequencies was <0.01 in MrBayes, or the maxdiff was <0.3 (in most cases <0.1) in PhyloBayes. We calculated 50% majority-rule consensus trees by pooling sampled trees from each of the 2 runs after excluding the burn-in. We excluded 25% of trees in MrBayes (in all cases long after the posterior likelihood of the sampled trees had reached a plateau), whereas for the PhyloBayes analyses, we

calculated the burn-in case by case to minimize the maxdiff statistics, which informs about convergence. In all Bayesian and ML analyses, a discrete gamma distribution with 4 rate categories was used to model among-site rate variation. In the remaining part of the text, models will be referred without reference to the gamma parameter (which was always implemented).

Reanalyses

The R2010 nucleotide data set.—Bayesian and ML analyses of the nt data set were performed employing the same model used by R2010 (nt-GTR without codon partitioning). A Bayesian analysis of the nt data set was also performed using the nt-CAT model, and 10-fold Bayesian cross-validation (Stone 1974) was used to test whether nt-GTR or nt-CAT fits this data set better. For the cross-validation, we used a training set composed of 90% of the sites in the alignment and a test set composed of the remaining 10% of the sites. Sampling from the Markov Chain Monte Carlo (MCMC) chains was performed every 100 generations until a total of 1000 trees was collected. A posterior predictive χ^2 analysis of composition homogeneity (Blanquart and Lartillot 2008) was performed on the nt data set. For the posterior predictive analysis, a burn-in of 100 points was used, and a total of 2000 points were sampled, with a frequency of 1 in 10. Bayesian cross-validation and posterior predictive analyses were performed using PhyloBayes.

R2010 amino acid data set.—We followed R2010 and first performed Bayesian and ML analyses of the aa data set under JTT. We then analyzed the aa data set using more realistic models of amino acid substitution whose parameters were inferred directly from the data. Bayesian and ML analyses were performed under an aa-GTR model, and Bayesian analyses were performed under the site-heterogeneous aa-CAT and aa-CATGTR models. Bayesian cross-validation (see nt analyses for settings) was used to test the fit of the 4 considered models of amino acid evolution (JTT, aa-GTR, aa-CAT, and aa-CATGTR). Finally, Bayesian analyses of a Dayhoff-recoded version of the aa data set were performed using the CAT and the CATGTR model.

Exploration of the Signal in the nt and aa Data Sets

Use of the slow–fast method.—To explore the nature of the signal in the nt and aa data sets, we excluded fast evolving sites using a slow–fast-based approach (Brinkmann and Phillipe 1999). The evolutionary rate of each site in the aa and nt alignments was estimated as its cumulative Parsimony score (*P*-score). This is obtained by summing, for each character, the *P*-scores calculated on 6 widely recognized groups (Arachnida, Malacostraca, Hexapoda, Branchiopoda, Myriapoda, and outgroup). Sites in the nt and aa alignments were then ranked according to their cumulative *P*-scores. We analyzed multiple data sets, each generated by subsequently excluding sites corresponding to a certain range of cumulative *P*-score values, starting with the fastest evolving sites (those with highest cumulative *P*-score). We sequentially excluded variable amounts of sites from the initial alignment in order to define sub-alignments corresponding approximately to a sequential reduction of 5% in size. The nt sub-data sets were analyzed with ML under nt-GTR, whereas the aa sub-data sets were analyzed with Bayesian analysis under aa-CAT.

Alternative amino acid character recoding using CAT.—To distinguish alternative codon families for arginine, leucine, and serine in aa-level analyses, we introduced new recoding strategies allowing the number of possible character states in the data set to be 21 or 23. By defining 21 states, we could discriminate between the 2 serine (TCN and AGY), leucine (CTN and TTR), or arginine (CGN and AGR) codon families. These recoding strategies were named, respectively, aa21S, aa21L, and aa21R. By defining a 23-state recoding strategy (named aa23LRS), we were able to simultaneously account for all 3 synonymous codon families. The CAT model implemented in Phylobayes has been adapted to assume an infinite mixture of equilibrium frequency profiles defined on vectors of 21 or 23 states. The aa21 recoding strategies showed convergence problems, as 1 out of 3 different chains recovered a different topology, resulting in a maxdiff = 1. Analysis using aa23LRS satisfactorily converged. The 21- and 23-state CAT models will be implemented in the forthcoming version of PhyloBayes.

Exclusion of serine synonymous codon sites from the nucleotide data set.—Starting from the R2010 "codon" data set, we generated a new nt data set in which all serine codons (both TCN and AGY) were recoded as missing. From this data set, hereafter referred to as missingSnoLRall1 + nt2, we then excluded third-codon positions as well as both the noRLall1 sites, and other masking characters that were excluded in R2010. This resulted in a data set identical to noLRall1 + nt2 but that was serine-free. The data sets were analyzed using both nt-CAT (the best fitting model—see above) and nt-GTR.

Codon-usage bias, skew, and compositional analyses.—Codon usage has been estimated using GCUA (McInerney 1998). We quantified taxon-specific biases in synonymous, serine codon families (TCN or AGY) usage, with a statistic based on skew (Perna and Kocher 1995). This statistic, named TCN/AGY skew, is calculated as (TCN - AGY)/(TCN + AGY) and ranges from +1 to −1. If TCN/AGY = 1 only TCN codons are used. If TCN/AGY = −1 only AGY codons are used. If TCN/AGY = 0 both codon types are equally used. AGY and TCN codons differ because of the presence or absence of A and G nucleotides. Accordingly, for each taxon in the data set, we plotted the percentage of A+G against its TCN/AGY skew. Similar skew values have been calculated for the leucine and the arginine codons: the CTN/TTR

skew (leucine) and the CGN/AGR skew (arginine). For each taxon, we plotted the percentage of T against its leucine CTN/TTR skew, and the percentage of A against its arginine CGN/AGR skew. Frequencies for the 4 nucleotides were calculated using constant amino acid positions that are encoded by 4-fold degenerate sites: these sites can be considered to be approximately neutral and should provide an estimate of the mutational pressure acting on the genome. To display the serine codon skew, we colored lineages according to predefined skew classes, using intervals of 0.05 units. We started from a class of skew characterized by the highest bias toward TCN codons (colored in dark gray) and progressively used lighter colors for classes more biased toward AGY. For each major clade, we calculated the mean skew and its standard deviation, and colored the corresponding box (in the right part of the tree) using the same coloring scheme used for individual taxa. We also identified leucine and arginine skew values on the nucleotide tree using an interval unit of 0.2.

Results and Discussion

Model Improvements Do Not Solve the Incongruence of Nucleotide and Amino Acid-Based Phylogenies

According to a χ^2 test, NoLRall1 + nt2 (the nt data set we reanalyzed in this study) is the most compositionally homogenous data set of R2010. We further tested the homogeneity of this data set using a posterior predictive homogeneity test under the nt-CAT model, and the data set passed the test (PP = 0.11). The tree inferred using this data set strongly supports the 6 new pancrustacean groups of Regier et al. (2010)—the nodes with black circles in Fig. 1a. The tree inferred from the analysis of the corresponding amino acid translation (Fig. 1b under the JTT model of amino acid replacement) is, on the other hand, poorly resolved and displays only 3 of the 6 nt-supported, pancrustacean groups (Xenocarida, Communostraca, and Multicrustacea).

The nucleotide topology appears to be robust to the hypothesis of among-site homogeneity of the substitution process. That is, trees under either the site-heterogeneous nt-CAT or the site-homogeneous nt-GTR models are nearly identical (Fig. 1). Alternatively, the amino acid phylogeny is more model dependent: when the amino acid data set is analyzed using better fitting, site-heterogeneous models (aa-CAT and aa-CATGTR, see cross-validation scores in Supplementary Table S1), we recovered trees (Fig. 1c,d and Supplementary Fig. S1) that are in disagreement with both the *nt* tree of Figure 1a and the JTT-derived *aa* tree of Figure 1b. Of all the novel pancrustacean groups supported by the nt data, only Communostraca (i.e., Malacostraca plus Thecostraca) is recovered. In addition, and regardless of the position of the various crustacean lineages, Branchiopoda and Copepoda appear to be more closely related to Hexapoda and Remipedia than the Malacostraca are. Here, we will refer to the group composed of Branchiopoda, Copepoda, Hexapoda, and Remipedia as to Group A. Oligostraca plus Communostraca will be referred to as Group B. The pancrustacean relationships reported in Figure 1c,d (displaying Groups A and B) are in agreement with recently published aa-based phylogenomic studies that employed more genes, but more restricted sets of taxa (Meusemann et al. 2010; Campbell et al. 2011; Von Reumont et al. 2012). An important aspect of the trees in Figure 1c,d is that they provide support for a recognized clade within Arthropoda (Chelicerata) for which the nucleotide and JTT trees of Figure 1a,b found weak to no support. Overall, the signal in the aa data set for the relationships within Pancrustacea is weak (Fig. 1b–d), and low support values were expected. However, some of the groups supported by the nt data set (in particular Vericrustacea) receive extremely low support (PP < 0.01) from the aa data set, under the best-fit aa model. That is, the signal supporting Vericrustacea is essentially absent from the aa data set. This reveals a clear difference between data types and calls for an in-depth analysis of their signal.

Subtle Phylogenetic Signal in Both nt and aa Data Types

What is causing the differences observed when comparing the aa and nt trees? An interesting clue comes from a slow–fast experiment (Brinkmann and Phillipe 1999), showing that signal supporting some of the pancrustacean subgroups of Figure 1a is concentrated in the fastest evolving sites (Supplementary Fig. S2a). In particular, signal for Vericrustacea is seriously reduced after exclusion of approximately fastest evolving 15% of the sites (Supplementary Fig. S2b). Fast-evolving sites are a potential source of misleading information because, being less constrained, they are the most likely to be saturated (see "Nucleotides Versus Amino Acids" section) or affected by compositional biases (Rodríguez-Ezpeleta et al. 2007). This suggests that some groups in the nt tree of Figure 1a, particularly Vericrustacea, may represent tree-reconstruction artifacts. The same slow–fast experiment using the aa data shows that signal in this data type is more stable when analyzed using the aa-CAT model. Signal for both Group A and particularly Group B is the prevailing one regardless of exclusion of various classes of fast-evolving sites (Supplementary Fig. S2c). Overall, the slow–fast experiments reveal that competing signals are present in the nt alignment, whereas model dependency of the aa topology suggests that genuine phylogenetic signal is weak in the aa alignment.

Incongruence Between nt- and aa-Based Phylogenies Is Caused by Synonymous Codon Families

The nt topology displays 6 new pancrustacean groups, including Vericrustacea and Altocrustacea (Fig. 1a), whereas the aa tree displays Group B and, less congruently, Group A (Fig. 1c,d). One possible

explanation for such discrepancies would be that amino acids are unable to discriminate between serine codons of the TCN versus AGY families. Because 2 non-synonymous transversions are needed to transform one serine codon into another one, such synonymous codons may contain reliable phylogenetic signal. Under this interpretation, the nucleotide tree (Fig. 1a) would be fundamentally correct, and the lack of resolution for the pancrustaceans in the amino acid trees (Fig. 1b–d) would be the consequence of having ignored the phylogenetically informative characters represented by alternative serine codons. This argument predicts a loss of support for some of the groups (due to a loss of signal—see also R2010), but it does not explain the Bayesian support *against* Vericrustacea (PP < 0.01) in the site-heterogeneous aa analyses. Furthermore, support for recognized groups such as the Chelicerata is poor in both the nt tree of Figure 1a and the JTT aa tree of Figure 1b, whereas support for Chelicerata is high in aa trees inferred using site-heterogeneous models (Fig. 1c,d). This points toward a model misspecification problem, rather than to a general lack of signal caused by the inability to detect substitutions between the 2 classes of serine codons.

We investigated the extent to which serine codons can explain the differences between the nt and aa trees. In the R2010 data set, 48 528 cells are occupied by serines; of these, 33 679 are coded by TCN codons and 14 849 by AGY codons. A total of 2204 alignment positions contain at least 2 taxa with the amino acid serine. Interestingly, most of the replacements between the 2 serine codon families are concentrated in Pancrustacea (Supplementary Fig. S3). Analyses of the missingSnoLRall1 + nt2 data set (where serine is recoded as missing) performed under GTR results in a drastic drop in support for the crustacean clades specific to the nt tree (Vericrustacea BS falls from 87 to 45; Xenocarida BS falls from 98 to 31; and Altocrustacea from 96 to 44, cf. Figs. 1a and 2a). When the same experiment is repeated using the better fitting site-heterogeneous nt-CAT model, the nt topology (Fig. 2b) becomes very similar to that of the aa trees of Figure 1c,d, and Vericrustacea, Altocrustacea, and Xenocarida are not recovered anymore. These results confirm that the signal for most of the pancrustacean groups of Figure 1a is conveyed by a subset of synonymous substitutions concentrated in a small proportion of sites: those characterized by substitutions between synonymous serine codon families (TCN and AGY). Signal for other nodes such as Mandibulata, Arthropoda, and Communostraca, on the other hand, seems more homogeneously distributed and remains high (or even increase in the case of Chelicerata). Interestingly, when the signal associated with serine codons is excluded, the pancrustacean nt topology becomes model dependent: this can be explained by considering that the exclusion of serine sites allowed other hidden signals to emerge, or more likely, by a lack of phylogenetic signal for the pancrustacean relationships in the non-serine part of the data set.

The Effect of Synonymous Codons Confirmed by 21-/23-State aa-CAT Model

The key role of serine is further confirmed by using a new character recoding strategy distinguishing between the 2 synonymous serine codon families (TCN and AGY, thus resulting in a total of 21 states). Using the CAT model on this recoded data set, we recovered almost the same topology (Fig. 2d) as did the nucleotide data set noLRall1 + nt2 of Figure 1a, except for the monophyly of Chelicerata. A similar topology is obtained when the leucine synonymous codons (TCN and TTR) are distinguished, although support levels in this case are low (tree not shown). Distinguishing between arginine codons (CGN and AGR) converged on the same topology of the 20-amino acids CAT model (not shown). Results obtained using these 21-state recoding strategies are not, however, reliable because only 2 out of 3 independent runs converged on the same tree topology. This is mirrored by low support at nodes of interest. On the other hand, a 23-state recoding strategy that can concomitantly distinguish between serine, arginine, and leucine synonymous codon families recovers the nt topology with higher support (and 3 independent runs converged on the same topology; values in brackets in Fig. 2c). These experiments confirm that synonymous codons are mainly responsible for the discrepancy between data types.

Replacements Between TCN and AGY Serine Codons Are Not Uncommon

Is the signal restricted to synonymous codon families reliable? The answer rests on understanding the nature of these substitutions. Because synonymous substitutions are silent, they can easily be driven by mutational pressures or by lineage-specific codon-usage preferences. Accordingly, it is common practice to exclude third-codon positions and, more rarely, leucine and arginine codons when analyses are performed using nucleotides (see Regier et al. 2010). Serine codons, on the other hand, are more often retained, mainly because of the implied double transversions, a putatively rare event. There are 2 ways of moving between the 2 TCN/AGY serine codon types: 2 simultaneous transversions (Averof et al. 2000) or 2 consecutive transversions through an intermediate non-serine state. Inspection of the 2204 positions in the alignment at which at least 2 taxa bear serine reveals that more than half of those positions contain at least one other taxon with threonine or cysteine, accommodating the possibility that both replacement routes are used. Indeed, the more likely intermediate between the 2 serine codon families is ACN, which codes for threonine. A second, but less likely, intermediate between the 2 serine codon families is TGY, which codes for cysteine. Both serine and threonine are polar hydrophilic hydroxyl-bearing amino acids, with similar phosphorylation propensities, and similar steric effects (Supplementary Fig. S4a). Indeed, the replacements between serine and threonine are

FIGURE 2. Synonymous serine codons are responsible for data-type conflict: When signal associated with replacement between synonymous serine codons is excluded, support for most of the pancrustacean clades of Figure 1a disappears and a topology more consistent with those supported by the amino acids emerges. a) Bayesian and ML analyses of the missingSnoLRall1 + nt2 data set using the nt-GTR model; b) The same data set (missingSnoLRall1 + nt2) using the nt-CAT model; c) Bayesian analyses of the amino acid data set employing the aa21S and aa23LRS recoding strategies under the aa-CAT model. Support values at nodes in panel c) are posterior probabilities from the consensus of 3 independent Phylobayes analyses. Lineages collapsed for clarity, see the main text and Figure 1 legend for further details.

among the most likely to occur, this can be observed (Supplementary Fig. S4b) by inspecting any of the many empirical amino acid replacement matrices available (Le and Gascuel 2008; Rota-Stabelli et al. 2009) or by observing that the serine/threonine profile is well populated in the aa-CAT model analyses (Lartillot and Philippe 2004). Similarly, serine and cysteine have a similar steric effect, and the most accepted change for cysteine in replacement matrices (Supplementary Fig. S4a) is serine. Transitions between serine codons belonging to different families might be more common than generally assumed, as they can happen without major disruption of both protein structure and function when mediated by threonine and to a lesser extent cysteine.

Synonymous Codon Usage Is Biased and Related to a Nucleotide Composition Bias

If the transition between the 2 classes (TCN and AGY) of serine codon families is more frequent than generally assumed, then serine codon usage could easily suffer from a lineage-specific bias due, for instance, to tRNA imbalance, mutational pressure, or a combination of both phenomena. In particular, mutational pressures are a well-known source of systematic errors (e.g., in bacterial phylogenetics; Embley et al. 1993; Cummins and McInerney 2011).

We initially tested if the synonymous codon usage for serine (but also leucine and arginine) is homogeneously distributed among taxa. To measure synonymous codon usage, we used a statistic based on skew (Perna and Kocher 1995—see "Materials and Methods" section). We also developed a posterior predictive homogeneity test to identify compositional problems in codon usage. This test employs the skew as the test statistic, and clearly rejects homogeneity of leucine, arginine, and serine skews (posterior predictive p-value < 0.01 for all 3 amino acids and Z-scores = 159, 200, and 10, respectively), suggesting that codon usage for these amino acids is differently biased in different arthropod lineages.

We further investigated whether the codon-usage bias is related to a nucleotide composition bias. To do this, we plotted the skew values versus nucleotide composition calculated at 4-fold degenerate sites. The correlation between the TCN/AGY (serine) skew and the A+G content is not significant over Arthropoda (R = 0.1588, P = 0.0705 after correcting for phylogenetic correlation; Supplementary Fig. S5a, left). However, it is stronger, though marginally significant, in non-insect pancrustaceans (R = 0.3328, P = 0.0466 after correcting for phylogeny in Supplementary Fig. S5a, right), suggesting that mutational pressure may be related to the serine bias, but only in non-insect pancrustaceans, the groups for which incongruence between data types exists. Leucine and arginine codon usage is also correlated with nucleotide composition in all arthropods and in non-insect pancrustaceans (Supplementary Fig. S5b,c, respectively). The marginal significance and the restricted phylogenetic scope of these correlations indicate that a compositional bias is unlikely to be the only cause for the serine bias: other forces may drive the observed codon-usage bias, for example, differential representation of tRNAs in different pancrustacean lineages.

Clusters of Taxa With Similar Codon-Usage Bias in the Nucleotide Tree

We investigated patterns of distribution of codon-usage preference (skew values) on the nucleotide topology. Arthropod lineages use TCN serine codons with skew values ranging from 0.25 to 0.55 (Fig. 3a). Interestingly, Branchiopoda, Copepoda, and Malacostraca use more AGY codons (light gray) than other crustacean classes, whereas Oligostraca and non-pancrustacean taxa use fewer AGY codons (dark gray), a distribution roughly reflecting the nucleotide tree topology. We thus conjecture that the nt tree of Figure 1a might be explained as the result of an attraction between lineages with similar serine codon usage: Branchiopoda and Copepoda would be attracted by the AGY-rich Malacostraca, whereas Oligostraca would be attracted toward the TCN-richer non-pancrustacean lineages. This would ultimately explain why exclusion of the signal associated with the serine codons (Fig. 2a) causes a loss of support for almost all the pancrustacean groups of Figure 1a. To further confirm the validity of our conjecture, and the potentially biasing effect of lineage-specific codon-usage preferences, we sub-sampled taxa from the R2010 nt data set to eliminate TCN-rich lineages (particularly Oligostraca). This resulted in a new data set where most retained crustaceans are AGY-rich. This data set unsurprisingly supports a tree where the remaining crustacean lineages are grouped (or rather attracted) into a monophyletic Crustacea (Fig. 3b). The same taxon subsampling experiment, if performed using the aa data set, has no effect: Group A and Group B are still recovered (tree not shown) with PP = 0.99 and 0.97, respectively. The serine codon-usage bias would also explain why support for R2010's novel pancrustacean groups is stronger when 4-state nucleotide models are used rather than 61-state codon models. Four-state models do not recognize if a substitution is silent or not, whereas 61-state codon models do; codon models will still use information from silent substitutions for likelihood calculations, but the impact of these substitutions on the final likelihood of a tree will be down-weighted. Another cluster of similarly biased taxa (Copepoda and Communostraca) can be observed when mapping leucine codon usage on the nucleotide tree, while no evident correlation is noticed when repeating the experiment with arginine (Supplementary Fig. S6).

Serine Codon Usage: A Subtle Bias That May also Affect Amino Acids

Our results suggest that the nt tree of Figure 1a might be misled by a lineage-specific serine codon-usage bias.

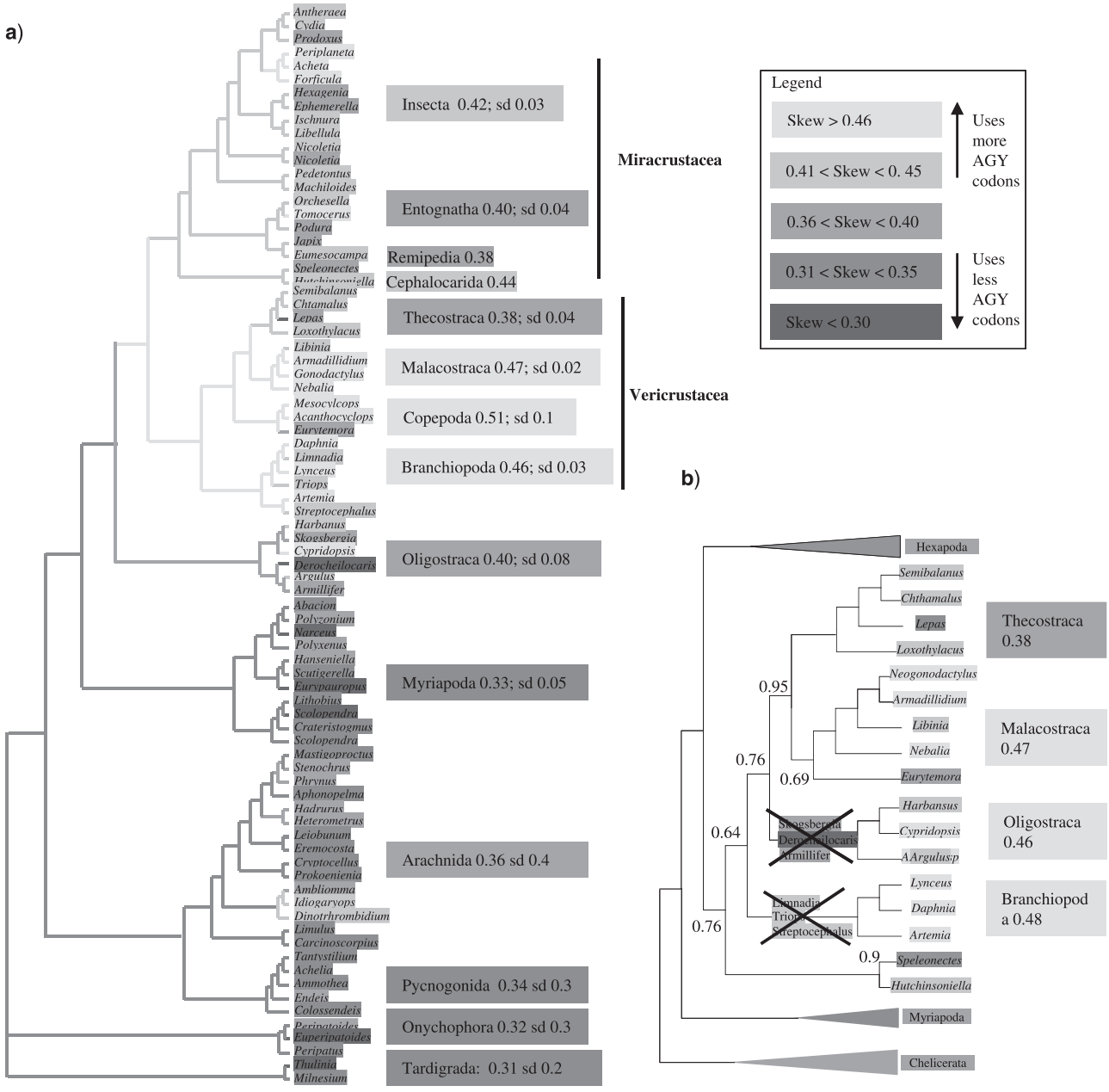

FIGURE 3. Serine codon-usage bias mapped on the nt tree: a) A map of TCN/AGY skews on the nt tree of Figure 1a (see main text for details). Species were color coded according to their skew (preference of TCN or AGY) as indicated in the legend, reported to the right of the tree. Species in light gray are those preferring AGY codons. All species are characterized by positive skew values, this means that they prefer TCN over AGY codons. For clarity, we sorted skew values in classes, and identify classes of skew using a scale of gray (see also main text and the legend on the left of the tree). A phylogenetic trend throughout the tree is clearly evident: non-pancrustacean lineages uses less AGY and are mainly colored using dark gray variants. Pancrustacean classes that use more AGY (Malacostraca, Copepoda, and Branchiopoda) are colored using lighter gray variants and are all grouped in the Vericrustacea. For each major clade in the figure, the average skew and its standard deviation are reported. b) Tree recovered from the analysis of a subsample of the taxa in Figure 3a. To create the sample used in Figure 3b TCN-rich (dark gray) group were excluded. This results in most crustaceans being AGY rich. Notably, these remaining groups are attracted into a monophyletic Crustacea, suggesting that codon-usage biases can drive the results of phylogenetic analyses.

This bias is, however, subtle and might have multiple driving forces. Its compositional component seems marginal, and it is thus unsurprising that the noLRall1 + nt2 passed a nucleotide posterior predictive test of compositional homogeneity (p = 0.11). We suggest that this is most likely because a compositional homogeneity test performed at the nucleotide level may not be adequate to identify a weak compositional bias acting at the codon level. Putatively affected serine sites are populated by either T/A (at first position) or C/G (at second position) given that the 2 serine codons are TCN and AGY. T/A- or G/C-populated sites, however, are

not exclusive to serine. Accordingly, when performing a compositional homogeneity test at the nucleotide level, the heterogeneity signature in the serine sites could be overwhelmed by that of other, similarly populated sites corresponding to non-serine codons.

A question that remains open is the validity of the competing aa topology that emerges under site-heterogenous models of aa evolution (Fig. 1c,d). As the amino acid models used are time-homogenous, the validity of the trees in Figure 1c,d depends, among other factors, on the presence of compositional biases at the level of the encoded amino acid sequences. We previously discussed how AGY and TCN substitutions may be mediated by threonine and (less likely) cysteine, which appear to be biochemically acceptable alternative amino acids. We further analyzed those positions in the alignments at which at least 2 taxa have serine. When only one type of serine codon (either TCN or AGY) is present, the number of positions at which another taxon displays either threonine or cysteine (422 positions) is similar to those positions not displaying them (405). When both types of serine codons (TCN and AGY) are present, we observe an excess of positions with threonine- or cysteine-bearing taxa (878) compared with those without any threonine- or cysteine-bearing taxa (499). This evidence cannot exclude simultaneous substitution between TCN and AGY (Averof et al. 2000). However, it points toward the existence of a subtle interplay between nucleotide mutation pressure, codon-usage bias, and amino acid composition.

Because we cannot exclude that even the aa data set is free from compositional biases, we explored whether the aa-supported groups (particularly Groups A and B) were affected by compositional-related artifacts and propose a way of dealing with it. We initially plotted the first 2 components (jointly explaining 90% of the compositional variation) of a principal component analysis of the 20 amino acid frequencies (Fig. 4a). We observed a tendency of Branchiopoda and Hexapoda, 2 clades belonging to Group A in the aa trees, toward the negative second component values (circled in Fig. 4a). To test for a possible compositional attraction between Branchiopoda and Hexapoda, we then used the Dayhoff recoding strategy, which recodes the 20 amino acids into 6 groups on the basis of their chemical and physical properties. This approach excludes (frequent) replacements between similar amino acids (Hrdy et al. 2004) and reduces the effects of saturation and compositional bias. Notably, the 6 categories in the Dayhoff recoding scheme do not distinguish between serines and threonines, both of which are represented by the same state. The same recoding strategy treats cysteine as an independent character state. Accordingly, Dayhoff recoding can partially (only in the case of threonine) safeguard against possible altered aa occurrences caused by an underlying serine codon-usage bias. Results of the analyses of a Dayhoff-recoded data set performed under the CAT and the CAT-GTR models (Fig. 4b) are consistent with the results obtained from the analyses of the 20 states aa data set in supporting Group A and Group B. However, support values are weaker in the Dayhoff-recoded trees: Group B is poorly supported and internal branches of Group A collapse. This can be partially explained by the recoding strategy being responsible for the exclusion of (genuine) phylogenetic signal. More likely, this recoding strategy has excluded some unreliable signal (including some signal associated with the serine codon-usage bias), and the partially unresolved topology of Figure 4b is probably the most conservative picture of pancrustacean relationships that we can obtain from this data set.

Conclusions

An Alternative View of Pancrustacean Evolution?

Overall, our results reveal that some of the nucleotide-based pancrustacean relationships of Figure 1a are unlikely to be correct. Nucleotides support a group of Branchiopoda, Malacostraca, Thecostraca, and Copepoda (the Vericrustacea in Fig. 1a), whereas amino acids points toward an alternative hypothesis (Fig. 1c and d). We have suggested that this discrepancy is caused by a serine synonymous codon-usage bias that may have promoted an artifactual attraction among unrelated taxa with similar serine codon-usage; Branchiopoda, Malacostraca, and Copepoda were attracted into a likely artifactual Vericrustacea, whereas an attraction between Oligostraca and the outgroups resulted in an artifactual Altocrustacea (Vericrustacea plus Miracrustacea).

Our results suggest that the Pancrustacea aa topology is potentially more trustworthy than the nt topology, as it seems to escape some of the biases associated with synonymous codon families. Most of our aa analyses (Figs. 1c and 4b) point toward a group composed of Branchiopoda, Remipedia, Copepoda, and Hexapoda (Group A). Cephalocarida is also included in this group when the best fitting aa-CATGTR model (Fig. 1d) is used. Group A (regardless of the inclusion of Cephalocarida) is always recovered as the sister of a group composed of Malacostraca, Oligostraca, and Thecostraca (Group B). The same groups are recovered after excluding synonymous codon families from the nt data set, that is, when missingSnoLRall1 + nt2 is analyzed under the best-fitting nt-CAT model (Fig. 2a). These results are in better agreement with recently published results of EST data sets (Meusemann et al. 2010; Rota-Stabelli et al. 2011; Von Reumont et al. 2012). However, the aa tree is partially model dependent (cf. Fig. 1c,d), particularly with reference to the position of the Cephalocarida. In addition, given that the serine codon-usage bias has the potential to affect also aa data sets (see above), it cannot be excluded that these results, as well as those of Von Reumont et al. (2012), Rota-Stabelli et al. (2011), and Meusemann et al. (2010) are also affected by this bias to some extent. Future work should investigate this possibility. More fundamentally, it seems that despite the important sampling effort of R2010 and of other

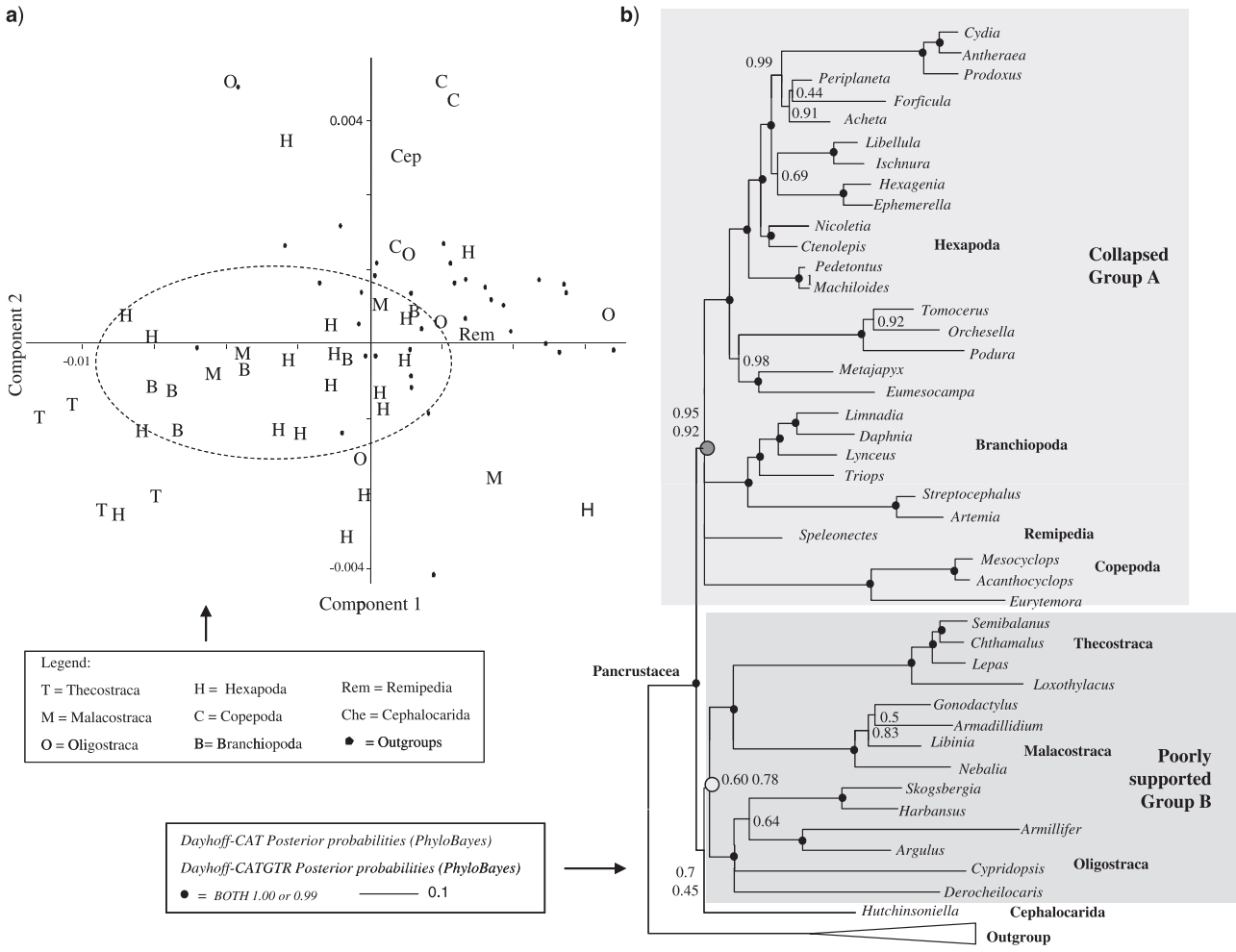

FIGURE 4. Possible compositional problems in the aa data set and Dayhoff corrections. a) Principal component analysis of the 20 amino acid frequencies. This panel shows a possible compositional attraction between Branchiopoda and Hexapoda. b) Dayhoff recoding analyses, designed to reduce the effect of compositional artifacts, recover a topology in accordance with that obtained using standard 20 aa state models (Fig. 1c). Overall, the results presented here are consistent with the possibility that the serine codon-usage bias we observed at the nucleotide level translates into an amino acid composition bias, even if they do not confirm it. Moderate support for Group A is recovered but its internal relationships are clearly unresolved (PP < 0.50). Group B is scarcely supported. See text for more details.

studies (Rota-Stabelli et al. 2011; Von Reumont et al. 2012): (1) more genes will be needed to resolve the pancrustacean relationships; (2) more lineages will have to be sampled, particularly to shorten long uninterrupted branches from the underrepresented Cephalocarida and Remipedia; and (3) detailed analyses will need to be performed to investigate whether other data sets are affected by the pancrustacean-specific serine codon-usage bias we identified.

Amino acids or nucleotides?—Our results provide an insight into the question of data-type choice in phylogenetics. We have shown that, with reference to this data set, the nt-based topology is affected by convergent replacements between synonymous codon families. This happens even in the serine case, where 2 non-synonymous transversions are involved. Compared with nucleotides, amino acids are more robust to compositional effects because, at the minimum, they will escape artifacts related to most synonymous substitutions. On the other hand, amino acids can still suffer from a general mutational pressure acting at the nucleotide level, hence the need to carefully perform extensive testing of potential compositional attractions with both data types (Foster and Hickey 1997; (Blanquart and Lartillot, 2008)). Because, at the least, some of the substitutions between serine codons might proceed through a threonine (or, less likely, a cysteine) intermediate, subtle interactions between codon usage and amino acid composition are possible. These interactions deserve careful future investigations.

Our cautious final interpretation is that, given the current evidence, amino acids should be preferred to nucleotides, even when the latter are analyzed at the codon level. This is in agreement with previous observations by (Inagaki and Roger, 2006) and Inagaki et al. (2004). When nucleotides are analyzed,

all synonymous substitutions should be removed. Alternatively, the use of non-stationary models (Foster 2004, Blanquart and Lartillot 2008) should be considered. In the long run, the use of codon models that can handle heterogeneity across sites (Rodrigue et al. 2010) should also be considered, but at this stage, the implementation of these models is still intractable.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.7p1k8304.

Funding

This work was supported by an individual Irish Research Council for Science, Engineering and Technology (IRCSET) Empower Fellowship, and a Marie Curie-Trento Province COFUND Fellowship [to O.R.S.]; and The Science Foundation Ireland Research Frontier Programme (SFI-RFP) [to D.P.] [Grants: SFI-RFP 08/RFP/EOB1595 and SFI-RFP 11/RFP/EOB/3106].

Acknowledgments

The authors thank James McInerney, Carla Cummins, Lahcen Campbell, David Fitzpatrick, Kevin Peterson, David Alvarez-Ponce, Peter Foster, Cymon Cox, and an anonymous reviewer for their useful comments and suggestions. All analyses were performed using the infrastructures of the Irish Center for High End Computing (ICHEC); and the National University of Ireland Maynooth High Performance Computing (HPC) and Réseau Québécois de Calcul de Haute Performance (RQCHP) facilities.

References

Altekar G., Dwarkadas S., Huelsenbeck J.P., Ronquist F. 2004. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415.

Averof M., Rokas A., Wolfe K.H., Sharp P.M. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science 287:1283–1286.

Blanquart S., Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842–858.

Brinkmann H., Phillipe H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Campbell L.I., Rota-Stabelli O., Edgecombe G.D., Marchioro T., Longhorn S.J., Telford M.J., Philippe H., Rebecchi L., Peterson K.J., Pisani D. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc. Natl Acad. Sci. USA 108(38):15920–15924.

Cook C.E., Yue Q., Akam M. 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proc. R. Soc. B 272:1295–1304.

Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60(6):833–844.

Embley T.M., Thomas R., Williams R. 1993. Reduced thermophilic bias in the 16S rDNA sequence from Thermus rubber provides further support for a relationship between Thermus and Deinococcus. Syst. Appl. Microbiol. 16:25–29.

Ertas B., von Reumont B.M., Wägele J.W., Misof B., Burmester T. 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. Mol. Biol. Evol. 26:2711–2718.

Fanenbruck M., Harzsch S., Wägele J.W. 2004. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. Proc. Natl Acad. Sci. USA 101: 3868–3873.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Foster P.G, Cox C.J., Embley T.M. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Phil. Trans. R. Soc. B 364:2197–2207.

Foster P.G., Hickey D.A. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284–290.

Foster P.G., Jermiin L.S., Hickey D.A. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44:282–288.

Galtier N., Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl Acad. of Sci. USA 92: 11317–11321.

Gibson A., Gowri-Shankar V., Higgs P.G., Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol. Biol. Evol. 22:251–264.

Holder M.T., Zwickl D.J., Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B 363:4013–4021.

Hrdy I., Hirt R.P., Dolezal P., Bardonova L., Foster P.G., Tachezy J., Embley T.M. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618–622.

Inagaki Y., Roger A.J. 2006. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Mol. Phylogenet. Evol. 40:428–434.

Inagaki Y., Simpson A., Dacks J., Roger A. 2004. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. Syst. Biol. 253:582–593.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Jermiin L.S., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.

Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Lanave C., Preparata G., Saccone C., Serio G. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Lockhart P.J., Howe C.J., Bryant D.A., Beanland T.J., Larkum A.W.D. 1992. Substitutional bias confounds inference of cyanel origins from sequence data. J. Mol. Evol. 34:153–162.

Mallatt J., Giribet G. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 40:772–794.

McInerney J.O. 1998. GCUA: general codon usage analysis. Bioinformatics 14:372–373.

Meusemann K., von Reumont B.M., Simon S., Roeding F., Strauss S., Kück P., Ebersberger I., Walzl M., Pass G., Breuers S., Achter V., von Haeseler A., Burmester T., Hadrys H., Wägele J.W., Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27:2451–2464.

Perna N.T., Kocher T.D. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. 41:353–358.

Regier J.C., Shultz J.W., Ganley A.R., Hussey A., Shi D., Ball B., Zwick A., Stajich J.E., Cummings M.P., Martin J.W., Cunningham C.W. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57:920–938.

Regier J.C., Shultz J.W., Kambic R.E. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. R. Soc. B 272:395–401.

Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083.

Regier J.C., Zwick A. 2011. Sources of signal in 62 protein-coding nuclear genes for higher-level phylogenetics of Arthropoda. PLoS One 6(8):e23408.

Rodrigue N., Philippe H., Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl Acad. Sci. USA 107:4629–4634.

Rodríguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Ronquist F., Huelsenbeck J.P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rota-Stabelli O., Campbell L.I., Brinkmann H., Edgecombe G.D., Longhorn S.J., Peterson K.J., Pisani D., Philippe H., Telford M.J. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. B 278:298–306.

Rota-Stabelli O., Kayal E., Gleeson D., Daub J., Boore J.L., Telford M.J., Pisani D., Blaxter M., Lavrov D.V. 2010. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol. Evol. 2:425–440.

Rota-Stabelli O., Yang Z., Telford M.J. 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol. Phylogenet. Evol. 52: 268–272.

Saccone C., De Giorgi C., Gissi C., Pesole G., Reyes A. 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene 238:195–209.

Seo T., Kishino H. 2009. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst. Biol. 58: 199–210.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B 36:111–147.

Von Reumont B.M., Meusemann K., Szucsich N.U., Dell'Ampio E., Gowri-Shankar V., Bartel D., Simon S., Letsch H.O., Stocsits R.R., Luan Y.X., Wägele J.W., Pass G., Hadrys H., Misof B. 2009. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evol. Biol. 9:119.

Von Reumont B.M., Jenner R.A., Wills M.A., Dell'Ampio E., Pass G., Ebersberger I., Meyer B., Koenemann S., Iliffe T.M., Stamatakis A., Niehuis O., Meusemann M., Misof B. 2012. Pancrustacean Phylogeny in the Light of New Phylogenomic Data: support for Remipedia as the Possible Sister Group of Hexapoda. Mol. Biol. Evol. 29:1031–1045.

Woese C.R., Achenbach L., Rouviere P., Mandelco L. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14:364–371.

Yang Z., Nielsen R., Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15:1600–1611.

Yang Z., Roberts D. 1995. On the use of nucleic acid sequences to infer early branches in the tree of life. Mol. Biol. Evol. 12: 451–458.