J Mol Evol (2002) 55:232–245
DOI: 10.1007/s00239-002-2321-5

JOURNAL OF MOLECULAR EVOLUTION

Gene Conversion and the Evolution of Euryarchaeal Chaperonins: A Maximum Likelihood-Based Method for Detecting Conflicting Phylogenetic Signals

John M. Archibald,[*] Andrew. J. Roger

Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7

Received: 27 November 2001 / Accepted: 26 February 2002

Abstract. Recombination is well known as a complicating factor in the interpretation of molecular phylogenies. Here we describe a maximum likelihood sliding window method based on a likelihood ratio test for scanning DNA sequence alignments for regions of incongruent phylogenetic signals, such as those influenced by recombination. Using this method, we identify several instances of gene conversion between paralogous chaperonin genes in euryarchaeote Archaea, many of which are not detected by two other widely used methods. In the *Thermococcus*/*Pyrococcus* lineage, where a gene duplication producing *a* and *b* paralogues predates the divergence of *Thermococcus* strains KS-1 and KS-8, gene conversion has homogenized portions of the *a* and *b* genes in KS-8 since the divergence of these two strains. A region near the 3′ end of the *a* and *b* paralogues in the methanogen *Methanobacterium thermoautotrophicum* also appears to have undergone gene conversion. We apply the method to two additional test data sets, the *argF* gene of *Neisseria* and a set of actin paralogues in maize, and show that it successfully identifies all the recombinant regions that were previously detected with other methods. Our approach is relatively insensitive to the presence of divergent sequences in the alignment, making it ideal for detecting recombination between both closely and distantly related genes.

Key words: Maximum likelihood — Recombination — Sliding window — Chaperonins — *argF* — Actin

Introduction

The advent of complete genome sequencing has facilitated large-scale comparisons of genomes from evolutionarily diverse organisms. Such comparisons have lead to the discovery that prokaryotic genomes are evolutionary chimeras, i.e., they are composed of genetic material from different sources, having acquired "foreign" genes through the process of lateral (or horizontal) gene transfer (Nelson et al. 1999; Ochman et al. 2000). It is becoming increasingly clear that genes themselves can also be chimeric in nature. In *Escherichia coli*, for example, mosaicism has been documented in the *gnd* (Bisercic et al. 1991; Dykhuizen and Green 1991) and *phoA* (DuBose et al. 1988) genes, and extensive intragenic recombination has also been observed within the genus *Neisseria*. The *IgA* protease gene of *N. gonorrhoeae* (Halter et al. 1989) and the *N. meningitidis argF* gene (Zhou and Spratt 1992) clearly possess mosaic structures, and the adenylate kinase (*adk*) and 16S ribosomal RNA genes of *Neisseria* species also appear to have been influenced by recombination (Feil et al. 1996; Smith et al. 1999). Interspecific recombination not only poses a challenge to the concept of a bacterial species, it can be of major biological significance. For example, interspecific, intragenic recombination in the penicillin-binding protein 2 (*penA*) genes of *N. gon-*

[*] *Present address:* Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada.

Correspondence to: Andrew J. Roger; *email*: aroger@is.dal.ca

orrhoeae, *N. meningitidis*, and *N. lactamica* has been important in the evolution of penicillin resistance (Smith et al. 1991; Spratt 1988; Spratt et al. 1989, 1992).

Gene conversion is a special form of recombination involving the non-reciprocal exchange of genetic information between paralogous genes within a genome. The process is well known as an important force in the evolution of multigene families, mediating, for example, the concerted evolution of ribosomal RNA (rRNA) genes in Eubacteria, Archaea, and Eukaryotes (Gangloff et al. 1996; Liao 2000; Scott et al. 1984). Gene conversion can also result in mosaic gene structures: this has been demonstrated in the β-chain constant-region genes of the mouse T-cell receptor (Rudikoff et al. 1992), the major histocompatibility complex multigene family in human (Hughes 1995), and the actin paralogues of maize (Moniz de Sá and Drouin 1996).

An assumption of most methods used to infer phylogenetic trees from molecular data is that all regions of a given gene have the same underlying evolutionary history. However, intragenic recombination leads to a situation in which this assumption is violated. In the interest of correctly inferring the history of genes, it is thus important to be able to detect when and where recombination events have occurred. To this end, several methods for detecting recombination have been developed and are diverse in terms of their underlying approaches. The methods of Stephens (1985) and Sawyer (1989) attempt to determine whether the distribution of nucleotide substitutions observed in an alignment of DNA sequences differs from a random distribution. McGuire et al. (1997) and Grassly and Holmes (1997) have developed sliding window-based methods for detecting recombination within a phylogenetic context, using the distance/least-squares and maximum likelihood methods of phylogenetic inference, respectively. Other phylogenetic methods include the "phylogenetic scanning" method of Fitch and Goodman (1991), the parsimony-based approach of Hein (1993), and a Bayesian method devised by McGuire et al. (2000). Many of these and other methods have been reviewed extensively elsewhere (Drouin et al. 1999; Weiller 1998; Wiuf et al. 2001).

Huelsenbeck and Bull (1996) described a likelihood ratio test for determining whether tree topologies estimated from different genes (for the same set of taxa) differ significantly from one another. Here we present a maximum likelihood method, based on a likelihood ratio test, for scanning nucleotide sequence alignments for intragenic regions of conflicting phylogenetic signal. Although the likelihood-based method of Grassly and Holmes (1997) is designed to detect regions of low average likelihood (given a full data set phylogeny) caused by recombination or positive selection, it does not distinguish between these two phenomena. In contrast, our method uncouples these factors and identifies only regions with phylogenetically discordant signals, indicating recombination. Our method uses a sliding window approach to identify regions of the alignment with phylogenies that are inconsistent with the phylogeny inferred from the molecule as a whole. We demonstrate its utility by focusing primarily on the evolution of archaeal chaperonins, a highly paralogous gene family in which gene conversion has recently been shown to be a significant factor (Archibald and Roger 2002). To test the performance of the method, we also examine the actin paralogues of maize and the *Neisseria argF* genes, gene families shown by others to have been influenced by gene conversion and recombination (Drouin et al. 1999; Grassly and Holmes 1997; Moniz de Sá and Drouin 1996; Zhou and Spratt 1992). The method proved useful for identifying regions of the alignments that were inconsistent with the majority phylogenetic signal, and examination of the phylogenies of such regions allowed the nature of the conflicts to be elucidated.

Materials and Methods

DNA Sequence Alignments

From a master alignment of archaeal chaperonin genes (Archibald et al. 1999), smaller alignments containing subsets of sequences were constructed. The alignment of Euryarchaeotes used for phylogenetic analysis contained 22 sequences and 1482 unambiguously aligned nucleotide positions. An alignment of chaperonin genes from the *Thermococcus*/*Pyrococcus* clade consisted of six sequences and 1554 positions. For methanogens, the alignment contained five sequences and 1557 sites, as did the alignment of four *Thermoplasma* sequences (α and β paralogues from *T. acidophilum* and *T. volcanium*). Finally, an alignment of five sequences from the halophiles *Haloferax volcanii* and *Halobacterium* NRC-1 contained 1530 unambiguously aligned sites.

A DNA alignment of a large diversity of plant actin genes was obtained by anonymous FTP from G. Drouin (University of Ottawa, Canada) at ftp://bio01.bio.uottawa.ca/pub/actin. A smaller alignment of eight *Zea mays* paralogues [Mac1 (J01238), Maz56 (U60514), Maz63 (U60513), Maz81 (U60511), Maz83 (U60510), Maz87 (U60509), Maz89 (U60508), and Maz95 (U60507)] was constructed from this data set and contained 975 unambiguously aligned positions. Finally, a partial sequence alignment containing a large diversity of *Neisseria argF* gene fragments was provided by E. Holmes (University of Oxford, England). This was used as a template to construct an alignment of eight full-length *argF* sequences (787 sites) from the following *Neisseria* species: *N. gonorrhoeae* FA19 (GenBank accession No. X64860), *N. meningitidis* HF46 and HF116 (X64865 and X64866, respectively), *N. polysaccharea* 11858 (X64870), *N. lactamica* 10617 (X64871), *N. cinerea* LNP1646 (X64869), *N. mucosa* LNP405 (X64873), and *N. flavescens* LNP444 (X64872). All alignments are available from A.J.R upon request.

Phylogeny

Phylogenetic analyses were performed using PAUP* version 4.0b8 (Swofford 1998). Maximum likelihood (ML) and ML-distance trees were inferred with a general time reversible plus Γ plus invariable sites (GTR + Γ+ P_{INV}) model using the heuristic search option. Starting trees for TBR branch swapping were obtained by the neighbor-joining method and distance trees were selected based on the minimum evolution criterion. A Γ distribution was approximated by four rate categories and the Γ shape parameter α, the proportion of invariable sites parameter, and the base frequencies were estimated from the data in PAUP*. Support for ML and ML-distance trees was obtained by bootstrapping with 100 resampling replicates.

Sliding Window Analyses

To identify discrete regions of alignments containing anomalous phylogenetic signal, we employed a ML sliding window method in which the ML tree obtained from the full alignment was compared to trees inferred from subsets of the data. Using "character set definitions" in PAUP*, a 100-nucleotide window was systematically advanced across the alignment in 10-nucleotide increments. For each window, the log-likelihood of the best tree from a heuristic ML search was obtained, as was the log-likelihood of the data present in the window given the *a priori* phylogenetic hypothesis, i.e., the ML tree obtained from the analysis of the complete alignment. The difference between these two log-likelihoods ($\Delta\ln L$) reflects the degree of conflict between the phylogenetic signal present in a given window and that in the molecule as a whole. A faster approximation was also performed in which log-likelihoods were inferred from ML-distance (minimum evolution) trees. $\Delta\ln L$ values were analyzed and plotted using Microsoft Excel.

To determine the significance of the $\Delta\ln L$ profiles, Seq-Gen version 1.2.4 (Rambaut and Grassly 1997) was used to simulate nucleotide sequence evolution over user-defined trees with the GTR + Γ + P_{INV} (= REV + Γ + P_{INV}) model. A null distribution was determined by simulating 500 data sets over the ML (or ML-distance) topology inferred from the full alignment and performing a sliding window analysis on each. The largest $\Delta\ln L$ value from each simulated data set was taken to form a distribution of maximal $\Delta\ln L$ values under the null hypothesis of no recombination with which empirical values were compared. For 500 simulations, the sixth highest $\Delta\ln L$ value is an estimate of the 0.99th quantile of the null distribution, and $\Delta\ln L$ values from the observed data greater than this value were considered significant at $p < 0.01$. This form of parametric bootstrapping is used because, in our method, the ML sliding window analyses are completed first, and regions of the alignment producing high $\Delta\ln L$ values are selected *a posteriori* as potentially discordant areas. Thus, to generate an appropriate null distribution, an *a posteriori* selection procedure must also be applied during parametric bootstrapping by selecting the highest $\Delta\ln L$ value ($\Delta\ln L_{\max}$) from each bootstrap replicate. A similar parametric bootstrapping procedure was described for the least-squares method of McGuire and Wright (2000).

Discrete regions of alignments identified as containing incongruent phylogenetic signal were investigated with additional phylogenetic analyses. The boundaries of such regions were taken as the first nucleotide of the first window and the last nucleotide of the last window possessing a $\Delta\ln L$ value greater than that of the null distribution.

A set of Perl scripts were written to implement the maximum likelihood sliding window (LIKEWIND) method described above (available at http://hades.biochem.dal.ca/Rogerlab/Software/software.html). *Likewind.pl* generates the PAUP* commands block used to perform the ML or ML-distance sliding window analyses, *getlikes.pl* obtains the $\Delta\ln L$ values from the *likewind.pl* outfiles, and *simblock.pl* produces the commands block containing the simulated data sets used in the determination of null distributions.

Additional Tests for Gene Conversion and Recombination

Two other programs were used to scan alignments for regions of incongruent phylogenetic signal. GENECONV (http://lado.wustl.edu-/~sawyer/geneconv/index.html) was used to perform Sawyer's (1989) statistical tests for detecting gene conversion. Briefly, Sawyer's method detects regions between pairs of sequences in an alignment that share more consecutive identical silent polymorphisms than would otherwise be expected by chance. For protein coding genes, it is possible that a given pair of sequences in an alignment exhibits significant similarity due to functional (selective) constraints on the protein sequences, rather than gene conversion. To control for this, GENECONV provides the option of focusing on "silent polymorphic sites"—degenerate sites in an alignment whose codons all specify the same amino acid (Drouin et al. 1999). Each of the alignments described above was used as input to GENECONV considering all polymorphic sites, as well as silent site-only (synonymous) polymorphisms (*-seqtype = silent*). Mismatch penalties of 0 (default) or 1 (*gscale* = 0–1) were used. N = 10,000 (default) random permutations of the polymorphic sites were performed in each analysis to assess the significance of putative gene conversion tracts.

We also employed the likelihood method of Grassly and Holmes (1997) using the program PLATO (http://evolve.zoo.ox.ac.uk/software.html). PLATO uses a sliding window approach to identify regions with "spatial phylogenetic variation." To do this, PLATO calculates a Q value that is defined as the average log-likelihood per site for a given window of the alignment divided by the average log-likelihood per site for the rest of the alignment. Log-likelihoods are calculated under the ML tree for the full data set. This Q value is calculated for all windows (≥5 positions but ≤50% of the alignment length) for all positions in the alignment. Maximum values of Q are associated with regions of low average site likelihood. Regions of significantly high Q are then evaluated by simulating the null distribution (parametric bootstrapping) of Q and flagging windows with observed values that fall above the null distribution as containing "spatial phylogenetic variation" (SPV). Parts of the alignment which exhibit SPV either possess a topology different from that of the null phylogenetic hypothesis (indicating that recombination has occurred) or have undergone changes in relative branch lengths or positive selection. The latter two explanations suggest that different branch lengths or Markov model parameters are optimal for the region, but not a different tree. Thus, PLATO detects model violations and/or recombination but does not uniquely identify which of these phenomena have occurred.

PLATO was run on each of the six data sets described above using the GTR (= REV) model of nucleotide substitution, taking into account among-site rate variation (ASRV) with an eight-rate category discrete approximation to the Γ distribution. The Γ shape parameter α and GTR rate matrix values were estimated from the data in PAUP*. Minimum sliding window sizes of 5, 50, and 100 nucleotides were used, and for each data set, the ML tree inferred from the complete alignment was used as the null hypothesis. Regions of our alignments that were identified by PLATO as anomalous but were not flagged as phylogenetically contradictory by LIKEWIND were tested for deviation from the optimal GTR + Γ + P_{INV} model for the whole data set by likelihood ratio tests. Briefly, the log-likelihood for the optimal phylogeny for the given window was maximized under the GTR + Γ + P_{INV} model with base frequency, rate matrix, rate variation, and branch-length parameters optimized for the window. The log-likelihood was then calculated for the window under the same tree with the GTR + Γ + P_{INV} model parameters optimized for the full data set (with branch lengths optimized for the window). As these are nested models, to determine whether the difference between these log-likelihoods ($\Delta\ln L$) is significant, $2\Delta\ln L$ can be compared to the

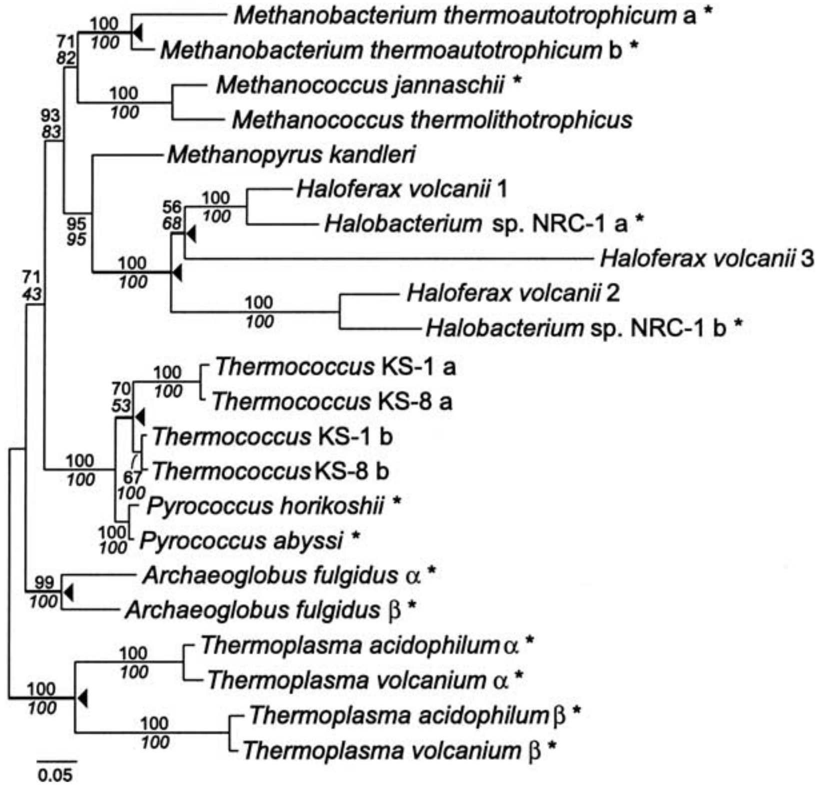
Methanobacterium thermoautotrophicum a *
Methanobacterium thermoautotrophicum b *
Methanococcus jannaschii *
Methanococcus thermolithotrophicus
Methanopyrus kandleri
Haloferax volcanii 1
Halobacterium sp. NRC-1 a *
Haloferax volcanii 3
Haloferax volcanii 2
Halobacterium sp. NRC-1 b *
Thermococcus KS-1 a
Thermococcus KS-8 a
Thermococcus KS-1 b
Thermococcus KS-8 b
Pyrococcus horikoshii *
Pyrococcus abyssi *
Archaeoglobus fulgidus α *
Archaeoglobus fulgidus β *
Thermoplasma acidophilum α *
Thermoplasma volcanium α *
Thermoplasma acidophilum β *
Thermoplasma volcanium β *
100
100
71
82
100
100
93
83
100
100
56
68
95
95
100
100
71
43
100
100
100
100
70
53
100
100
7
67
100
100
100
99
100
100
100
100
100
100
100
0.05

Fig. 1. Phylogeny of euryarchaeal chaperonins. The tree shown is a maximum likelihood (ML) tree (-ln*L*, 10,058.76) inferred from the first and second positions of an alignment containing 22 sequences and 1482 unambiguously aligned sites. The tree is arbitrarily rooted with the *Thermoplasma* sequences, consistent with previous analyses using amino acid sequences (Archibald et al. 1999, 200l). *Asterisks* appear next to sequences from *organisms* whose genomes have been completely sequenced, and inferred gene duplications are highlighted with an *arrowhead.* ML and ML-distance bootstrap values are provided for all branches in the tree (plain text and italics, respectively). The nomenclature used to identify the lineage-specific paralogues (e.g., *Haloferax volcanii* 1, 2, and 3, *Thermoplasma acidophilum* α and β) is consistent with their use in the literature. The scale bar represents the expected number of nucleotide substitutions per site.

χ^2 distribution with 10 degrees of freedom [degrees of freedom = free rate matrix parameters (5) + free base frequency parameters (3) + the Γ shape parameter α (1) + the proportion of invariable sites P_{INV} (1) = 10].

Results

Euryarchaeal Chaperonins

The chaperonins are a ubiquitous family of ATP-dependent molecular chaperone that form oligomeric double-ring complexes and mediate the folding of nascent proteins (for review see Bukau and Horwich 1998; Gutsche et al. 1999; Ranson et al. 1998). The evolution of archaeal chaperonins, which has been described extensively elsewhere (Archibald et al. 1999, 2001) is one in which gene duplication has played a prominent role. Archibald and Roger (2002) recently demonstrated that gene conversion has also been a factor. Gene conversions have occurred between two divergent crenarchaeal chaperonin paralogues independently in four lineages. Here we focus on identifying conversions between duplicate genes in euryarchaeal genomes.

A phylogenetic tree constructed from a DNA sequence alignment containing the full diversity of known euryarchaeal chaperonin genes is shown in Fig. 1. The most striking feature of the tree is the presence of numerous lineage-specific gene duplications. For example, in the *Thermococcus*/*Pyrococcus* clade, *Thermococcus* strains KS-1 and KS-8 each have two chaperonin paralogues, *a* and *b*, while two closely related *Pyrococcus* species (*P. horikoshii* and *P. abyssi*) each possess a single chaperonin gene. Phylogenetic analysis suggests that while the duplication producing the *Thermococcus a* and *b* genes predates the divergence of the KS-1 and KS-8 lineages, it occurred after the divergence of *Thermococcus* and *Pyrococcus* (Fig. 1). This is because the *a* and *b* paralogues branch together to the exclusion of the *Pyrococcus* genes. This result is in contrast to previous analyses performed using amino acid sequences which showed the *Pyrococcus* sequences branching specifically with the *Thermococcus b* paralogue (Archibald et al. 1999, 2001). Such a topology would suggest that the duplication producing the *a* and *b* paralogues of *Thermococcus* predates the *Thermococcus*/*Pyrococcus* split and that the *Pyrococcus* lineage has lost the *a* gene.

We used the LIKEWIND method described above to scan an alignment of *Thermococcus* and *Pyrococcus* chaperonin genes for regions of anomalous phylogenetic signal, such as those affected by gene conversion. The results are presented in Fig. 2A. When the ML topology inferred from the full alignment is used as the null hypothesis imposed on each of the windows, several Δln*L* peaks are readily apparent, two of which exceed the estimate of the 0.99th quantile of the null distribution (and are thus significant at $p < 0.01$). These peaks highlight regions of the

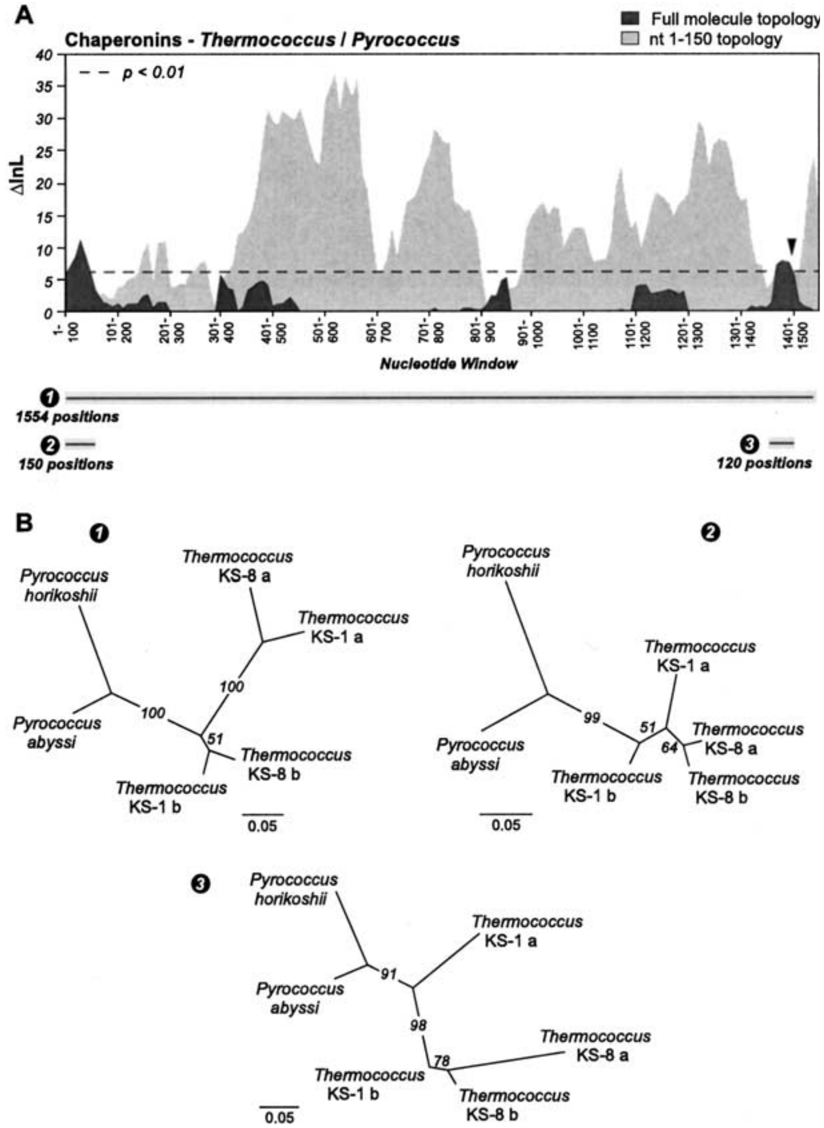

Fig. 2. Maximum likelihood (ML) sliding window ("LIKEWIND") analysis of chaperonin genes in *Thermococcus* and *Pyrococcus*. **A** Δln$L$ profiles obtained from the analysis of an alignment containing six *Thermococcus*/*Pyrococcus* sequences and 1554 unambiguously aligned nucleotide positions. The Δln$L$ profile obtained when the full molecule ML topology was used as the null phylogenetic hypothesis is plotted against that obtained when the tree constructed from nucleotide positions 1–150 was used. The *dashed line* indicates the estimate of the 0.99th quantile of the null distribution (Δln$L$ = 6.10), obtained by parametric bootstrapping with 500 replicates. The *arrowhead* highlights the approximate position of a three-nucleotide deletion present in the *a* and *b* paralogues of *Thermococcus* strain KS-8 (see Fig. 3). Three regions (data sets 1–3) of the alignment selected for phylogenetic analysis are also highlighted. Data set 1 contained the full alignment (1554 sites), while data sets 2 and 3 contained 150 and 120 positions, respectively. **B** ML trees inferred from the three data sets highlighted in A. Identical topologies were obtained using ML-distance. Branch support is provided (ML bootstrap values). Scale bars indicate the expected number of substitutions per site.

alignment that possess phylogenetic signal contradictory to that present in the molecule as a whole. To investigate this further, portions of the alignment corresponding to the two peaks were analyzed separately by phylogenetic analysis using ML and ML-distance methods. For reference, tree 1 in Fig. 2B shows the ML topology inferred from the complete alignment. This phylogeny clearly separates the *Thermococcus a* and *b* paralogues irrespective of organism and is identical to that obtained from analysis of the whole molecule minus the small regions producing the significant Δln$L$ peaks (data not shown). In contrast, phylogenies constructed from the 150-nucleotide region corresponding to the Δln$L$ peak at the 5′ end of the alignment show the *a* and *b* paralogues of *Thermococcus* KS-8 branching together (albeit weakly) to the exclusion of the other sequences (Fig. 2B, tree 2). This suggests that partial gene conversion has occurred between the *a* and the *b* paralogues in strain KS-8 in this region. Interestingly, the KS-1 *a* and *b* sequences are paraphyletic with respect to the other sequences in Fig. 2B, tree 2.

The region of the alignment corresponding to the Δln$L$ peak near the 3′ end of the gene is somewhat smaller than that at the 5′ end, approximately 120 nucleotides in length. Nevertheless, phylogenetic trees inferred from this region (Fig. 2B, tree 3) were also inconsistent with the "whole molecule" topology. As in tree 2 (Fig. 2B), the *a* and *b* paralogues of *Thermococcus* KS-8 branch together with reasonable bootstrap support, suggesting that an additional small-scale gene conversion has occurred in this region between the two sequences. A three-nucleotide deletion in the *Thermococcus* KS-8 *a* and *b* genes (Fig. 3) is also consistent with this interpretation. This deletion, whose distribution is at odds with the topology shown in Fig. 2B, tree 1, exists in a highly conserved region of the molecule, and is absent from the other *Thermococcus* and *Pyrococcus* sequences as well as all other archaeal genes (data not shown).

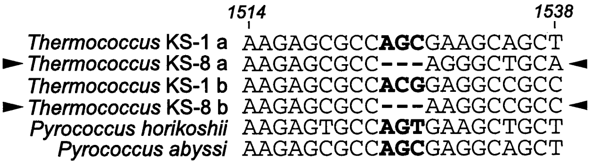
1514
1538
Thermococcus KS-1 a AAGAGCGCCAGCGAAGCAGCT
Thermococcus KS-8 a AAGAGCGCC---AGGGCTGCA
Thermococcus KS-1 b AAGAGCGCCACGGAGGCCGCC
Thermococcus KS-8 b AAGAGCGCC---AAGGCCGCC
Pyrococcus horikoshii AAGAGTGCCAGTGAAGCTGCT
Pyrococcus abyssi AAGAGCGCCAGCGAGGCAGCT

Fig. 3. A three-nucleotide deletion shared between the *a* and the *b* paralogues of *Thermococcus* strain KS-8. The approximate position of the deletion is highlighted in Fig. 2A. The coordinates provided are with respect to the complete *Thermococcus* KS-1 *a* gene sequence.

We also inferred a $\Delta\ln L$ profile from the *Thermococcus/Pyrococcus* data set using the alternate topology shown in Fig. 2B, tree 2, as the null phylogenetic hypothesis (Fig. 2A). As expected, most of the windows spanning the molecule produce $\Delta\ln L$ values well above the estimate of the 0.99th quantile, confirming the hypothesis that the bulk of the alignment contains phylogenetic signal that is strongly contradictory to that present in the first 150 nucleotides. Significantly, the $\Delta\ln L$ profile dips below the 0.99th quantile in most (but not all) of the regions corresponding to spikes in the "full molecule" profile (see Discussion).

To compare the ML sliding window method utilized above with existing methods for detecting recombination, the *Thermococcus/Pyrococcus* chaperonin data set was analyzed using Sawyer's (1989) method and the likelihood method of Grassly and Holmes (1997). Under the silent sites-only criterion, Sawyer's GENECONV program failed to detect the gene conversions identified above between the *a* and the *b* paralogues of *Thermococcus*, even allowing for gaps (see Materials and Methods). Using Grassly and Rambaut's PLATO program, two regions of the *Thermococcus/Pyrococcus* alignment were identified as being regions of significantly low likelihood, although, curiously, these regions did not correspond to those identified with the method described here. To investigate this further, we reanalyzed the *Thermococcus/Pyrococcus* data set in LIKEWIND without an invariable sites parameter (P_{INV}), such that the models of DNA sequence evolution used in the LIKEWIND and PLATO analyses were the same (i.e., GTR + Γ). The results were almost identical to those presented in Fig. 2A (data not shown), indicating that the difference in the performance of the two methods is not due to the use of slightly different model parameters. Likelihood ratio tests show that the regions identified by PLATO (but not LIKEWIND) correspond to areas which produced estimates of the parameters of the GTR + Γ + P_{INV} model significantly different from those inferred from the full alignment (e.g., positions 396–450, $\Delta\ln L = 11.31$, $p = 0.0042$; positions 1506–1554, $\Delta\ln L = 32.49$, $p < 1 \times 10^{-6}$). The results of the GENECONV, PLATO, and LIKEWIND experiments are summarized in Table 1.

Another euryarchaeal lineage in which chaperonin gene duplication has occurred is the methanogens. *Methanobacterium thermoautotrophicum* possesses an *a* and a *b* paralogue, unlike the other methanogens, *Methanopyrus kandleri*, *Methanococcus jannaschii*, and *M. thermolithotrophicus*, which each possess a single chaperonin gene [the presence of a single gene is confirmed in the case of *M. jannaschii*, whose genome has been completely sequenced (Bult et al. 1996)]. Phylogenetic analyses reveal that the duplication producing the *a* and *b* paralogues in *M. thermoautotrophicum* occurred after this organism diverged from the two *Methanococcus* species (Fig. 1) (Archibald et al. 1999, 2001).

The results of an ML sliding window analysis performed on the methanogen chaperonin alignment (Fig. 4A) suggest that the phylogenetic signal is relatively homogeneous across the full length of the molecule, as no regions of the $\Delta\ln L$ plot extend above the null distribution. To a certain extent, this is to be expected, given that the *a/b* gene duplication in *M. thermoautotrophicum* does not predate the divergence of this organism from other methanogens. Partial gene conversions would not result in an alternate tree topology, as the *a* and *b* paralogues are already more similar to each other than to any of the other sequences.

Gene conversions would, however, be expected to produce a situation in which the *a* and *b* paralogues appeared to be much more closely related to one another in discrete parts of the alignment. To test this hypothesis, a $\Delta\ln L$ profile was inferred under the "fixed-branch lengths" option, i.e., by evaluating the likelihood of the data in each window with the branch lengths of the null phylogenetic hypothesis imposed. Under this criterion, several regions of the $\Delta\ln L$ profile clearly exceeded the 0.99th quantile estimate (Fig. 4A). Most notably, a large peak near the 3′ end of the gene was identified, corresponding to nucleotides 1221–1480 (Table 1). Phylogenetic trees inferred from this portion of the alignment showed differences in branch lengths compared to those inferred from the rest of the molecule. The branches leading to the *M. thermoautotrophicum a* and *b* paralogues in tree 1, Fig. 4B, are much longer than those in tree 2, which was constructed from the putative gene conversion tract. Significantly, these results closely match those obtained with GENECONV under the silent site-only criterion (see Materials and Methods), which detected nucleotides 1207–1464 as an area of gene conversion between the *M. thermoautotrophicum a* and *b* paralogues (Table 1). PLATO did not detect this area as anomalously evolving, although it did flag several others (nt 792–805 and 1509–1533) as areas of low likelihood. As was the case for the regions identified by PLATO in the *Thermococcus/Pyrococcus* data set, the GTR + Γ + P_{INV} model parameters estimated

Table. 1. Results of GENECONV, PLATO, and LIKEWIND analyses using euryarchaeal chaperonin, *Neisseria argF*, and *Zea mays* actin alignments

Data set	GENECONV[a]	PLATO[b]	LIKEWIND[c]
Chaperonins			
Thermococcus/Pyrococcus	None detected	nt 396–450, $Z = 4.15$, nt 1506–1554, $Z = 7.96$	nt 1–150, 1371–1490
Methanogens	*M. thermoauto. a* & *b*, nt 1207–1464, $p = 6.64 \times 10^{-2}$	nt 792–805, $Z = 5.13$, nt 1509–1533, $Z = 14.06$	nt 791–890, 851–960, 1011–1120, 1221–1480[d]
Halophiles	None detected	nt 900–1007, $Z = 6.12$	nt 1311–1420
Thermoplasmas	*T.aci.* α & *T. vol.* β, nt 52–147, $p = 2.65 \times 10^{-2}$	nt 1528–1557, $Z = 8.96$	none detected
ArgF	None detected	None detected[e]	nt 1–220
Actin	*Z. mays* 56 & *Z. mays* 81[f] nt 847–975, $p = 7.23 \times 10^{-4}$	nt 870–972, $Z = 4.54$	nt 831–975

Note: M. thermoauto., Methanobacterium thermoautotrophicum; T. aci., Thermoplasma acidophilum; T. vol., Thermoplasma volcanium; Z. mays, Zea mays.

[a] Results are shown for analyses considering only silent-site polymorphisms and a gap penalty = 1 (see text).

[b] Results are shown only for analyses using a minimum sliding window size of 5 (default).

[c] Boundaries determined as described under Materials and Methods.

[d] Detected using the "fixed-branch lengths" option (see text).

[e] Using an HKY model without accounting for among-site rate variation, three regions near the 5′ end of the gene (nt 7–19, 60–64, and 171–181) were detected as regions of low likelihood, similar to that observed by Grassly and Holmes (1997).

[f] A longer gene conversion tract (nt 1–852) was also detected between the *Z. mays* 56 and the *Z. mays* 63 paralogues, although with marginal significance (see text).

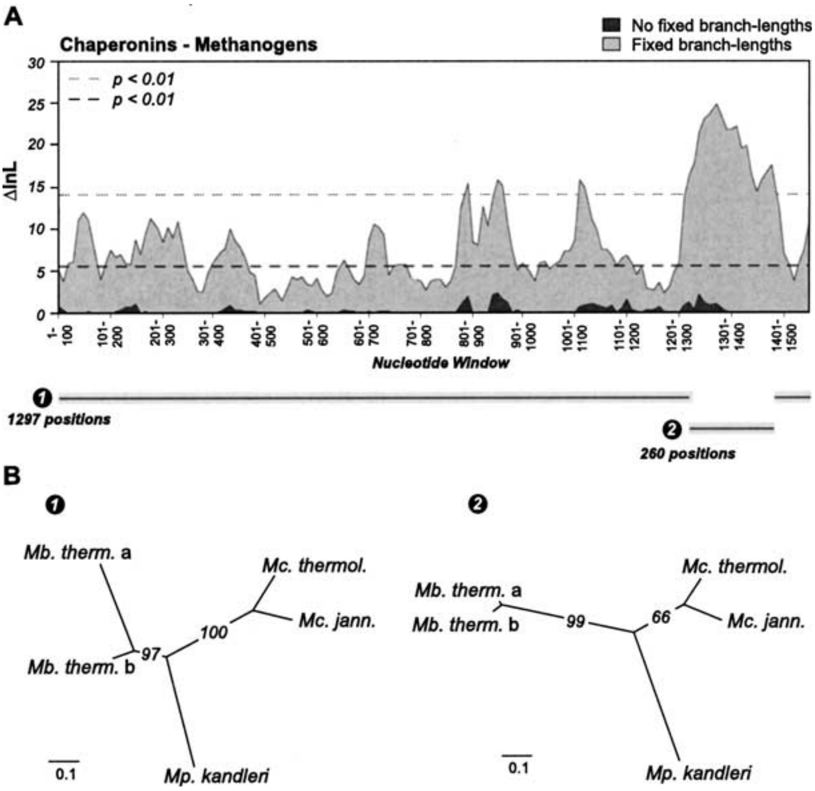

Fig. 4. Maximum likelihood (ML) sliding window analysis of chaperonin genes in methanogenic euryarchaeotes. **A** Δln*L* profiles inferred from an alignment of five sequences and 1557 nucleotide positions. Two profiles are plotted against one another (both obtained using the full ML topology as the null phylogenetic hypothesis), one inferred with the "fixed-branch lengths" option and the other without (see text). The *dashed lines* indicate the estimates of the 0.99th quantiles of their respective null distributions, obtained by parametric bootstrapping (Δln$L$ = 14.02 for the fixed-branch lengths plot, and Δln$L$ = 5.44 without fixed branch lengths). Two regions of the alignment selected for phylogenetic analysis are highlighted (data sets 1 and 2). **B** ML trees inferred from the two data sets in **A** (ML bootstrap values are provided). Scale bars indicate the expected number of substitutions per nucleotide position. *Mb. therm., Methanobacterium thermoautotrophicum; Mc. thermol., Methanococcus thermolithotrophicus; Mc. jann., Methanococcus jannaschii; Mp. kandleri, Methanopyrus kandleri.*

for nucleotides 1509–1533 were significantly different from those estimated from the complete alignment (Δln$L$ = 16.50, p = 0.00011).

Two additional areas of the euryarchaeal chaperonin tree in which lineage-specific paralogies exist were investigated for instances of recombination, using the same methods as above. Within the halophilic euryarchaeotes, a pair of gene duplications occurred in the common ancestor of *Haloferax volcanii* and *Halobacterium* sp. NRC-1. While *H. volcanii* has three chaperonin paralogues, the complete *Halobacterium* sp. NRC-1 genome (Ng et al. 2000) encodes only two chaperonin genes (*A* and *B*). The strong phylogenetic association of the *A* and *B* genes for *H. volcanii* paralogues 1 and 2 (Fig. 1) suggests that *Halobacterium* sp. NRC-1 at one time had, but subsequently lost, the third gene. Using LIKEWIND we identified a 120-nucleotide region (nt 1311–1420) of anomalous phylogenetic signal. Unlike the topology shown in Fig. 1, phylogenetic analysis of this region of the alignment yielded trees in which the *Halobacterium B* paralogue branched strongly with paralogue 3 of *Haloferax volcanii*, to the exclusion of the other sequences (data not shown). This suggests that intra- or intergenomic recombination has occurred between paralogue 3 and the *B* paralogue of *Halobacterium*. While GENECONV did not detect any gene conversion regions in the halophile data set, PLATO identified nucleotides 900–1007 as an anomalously evolving region. Phylogenetic analysis of this portion of the alignment produced trees with the same internal topology as shown in Fig. 1 (data not shown). Again, likelihood ratio tests confirmed that significantly different GTR + Γ + P_{INV} model parameters were estimated for this window versus those estimated from the whole alignment ($\Delta\ln L = 8.45$, $p = 0.023$). Analysis of an alignment containing the α and β paralogues of the two *Thermoplasma* lineages (*T. acidophilum* and *T. volcanium*) using our method produced no evidence for recombination. However, both GENECONV and PLATO identified small (although different) areas as possible recombinant regions (Table 1).

Neisseria argF

The *argF* gene encodes the protein ornithine transcarbamoylase, an enzyme in the arginine biosynthetic pathway. Zhou and Spratt (1992) showed that intragenic recombination has occurred in the *argF* gene of *Neisseria* species, and more recently, Grassly and Holmes (1997) analyzed the same data set using their PLATO program. Figure 5A shows the results of an ML sliding window analysis performed on an alignment containing eight *argF* sequences from a variety of *Neisseria* species and strains. The $\Delta\ln L$ profile reveals a large peak at the 5′ end of the gene. The phylogenetic tree inferred from the first 220 nucleotides of the alignment (Fig. 5B, tree 2) differs from the tree constructed from the whole molecule (Fig. 5B, tree 1) in that the *N. meningitidis* strains (HF46 and HF116) branch with *N. gonorrhoeae* in the "whole molecule" phylogeny, but with *N. cinerea* in the phylogeny constructed from the recombinant region. This is consistent with the results of Zhou and Spratt (1992) and and should Grassly and Holmes (1997). The bulk of the phylogenetic incongruence in the *argF* data set is due to the fact that the 5′ end of the *N. meningitidis* genes appear to be derived from a *N. cinerea*-like sequence.

We also inferred a $\Delta\ln L$, profile using the topology shown in Fig. 5B, tree 2, as the null hypothesis. As expected, elevated $\Delta\ln L$ values were observed for most regions of the alignment, except at the 5′ end (Fig. 5A), indicating that the signal present in most of the alignment is inconsistent with that present in the recombinant region. Interestingly, the area between nucleotide window 501–600 and nucleotide window 601–700 possesses elevated $\Delta\ln L$ values in both of the profiles shown in Fig. 5A. This area roughly corresponds to anomalous regions identified using Maynard–Smith's maximum χ^2 test (Zhou and Spratt 1992) and PLATO (Grassly and Holmes 1997) [nt 803–833 and 728–818, respectively, following the numbering system of Zhou and Spratt (1992)]. Phylogenetic analysis of this region yielded a third topology (tree 3, Fig. 5B) in which the *N. cinerea*, *N. mucosa,* and *N. lactamica* sequences formed a well-supported clade and the *N. gonorrhoeae* and *N. meningitidis* sequences did not cluster together. This indicates that recombination may also have occurred near the 3′ end of some of the *Neisseria argF* genes, although Grassly and Holmes (1997) suggest that this region is simply a highly polymorphic area.

Our analysis of the *Neisseria argF* data set using PLATO differed from the earlier study of Grassly and Holmes (1997) in taking into account among-site rate variation. When the GTR + Γ model was used, PLATO did not flag any regions of the *Neisseria argF* alignment as areas of significantly low likelihood. However, when a constant rates model was used (in our case, HKY with no rate heterogeneity), three short regions near the 5′ end of the gene were identified as having a low likelihood (Table 1), as observed previously (Grassly and Holmes 1997). These were contained within the recombinant region identified using LIKEWIND. GENECONV did not detect any regions of the alignment as being recombinant (Table 1).

Maize Actin

Drouin and colleagues have shown that gene conversion has been a factor in the evolution of the actin multigene family in the angiosperm *Zea mays* (Drouin et al. 1999; Moniz de Sá and Drouin 1996). These authors used Sawyer's method to identify two gene conversion tracts involving three different *Z. mays* paralogues. We analyzed an alignment of eight *Z. mays* actin genes using GENECONV, PLATO, and LIKEWIND.

The results of the ML sliding window analysis are shown in Fig. 6A. When the ML topology inferred

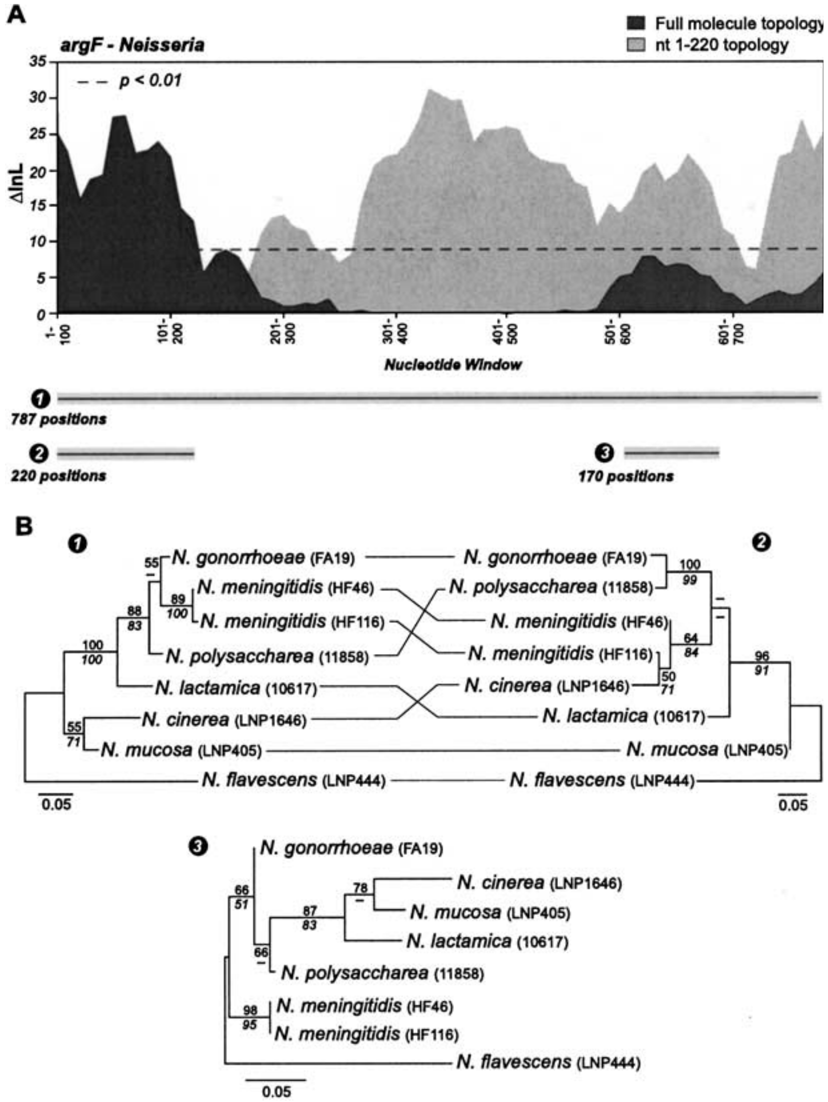

Fig. 5. Maximum likelihood (ML) sliding window analysis of *Neisseria argF* genes. **A** $\Delta\ln L$ profiles inferred from an alignment containing eight *Neisseria* sequences and 787 sites. Two $\Delta\ln L$ profiles are shown: one inferred with the full molecule topology as the null hypothesis, the other using the tree obtained from analysis of positions 1–120. The *dashed line* indicates the estimate of the 0.99th quantile of the null distribution ($\Delta\ln L = 8.77$), obtained by parametric bootstrapping. The approximate regions (and sizes) of three data sets (1–3) selected for phylogenetic analysis are shown. **B** Phylogenetic analyses of the three data sets highlighted in **A.** All three trees were constructed using the ML method. ML (plain text) and ML-distance (italics) bootstrap values are provided. Scale bars indicate the expected number of substitutions per nucleotide position.

from the complete molecule (the "full molecule topology") is used as the null phylogenetic hypothesis, the $\Delta\ln L$ profile is largely a flat line, with the exception of a single large peak at the 3′ end of the alignment. When the reciprocal experiment is performed, i.e., using the ML topology inferred from the 3′ end of the alignment (below) as the null hypothesis, large $\Delta\ln L$ values are obtained for all regions of the alignment except at the 3′ end. Phylogenies were inferred from three discrete portions of the alignment and are shown in Fig. 6B. The topology obtained from analysis of the complete alignment (tree 1) was the same as that inferred from the full alignment, minus 145 nucleotides at the 3′ end (tree 2). Significantly, these topologies were inconsistent with trees inferred from only the 3′ end of the alignment (tree 3). The three trees are identical except that the *Z. mays* 56 paralogue clusters strongly with paralogue 63 in trees 1 and 2, while branching with paralogues 81 and 83 in tree 3. These results are consistent with those of Moniz de Sá and Drouin (1996) and Drouin et al. (1999). GENECONV identifies nucleotides 847–975 (our coordinates) as a possible conversion tract between *Z. mays* 56 and *Z. mays* 81, or, alternatively, suggests that most of the coding region (nt 1–852) of paralogues 56 and 63 has been homogenized by gene conversion (Table 1). It is not clear which of these scenarios is correct. PLATO identifies a region near the 3′ end of the gene as having a significantly low likelihood (nt 870–972, $Z = 4.54$) but does not detect the area corresponding to the possible large-scale conversion event between paralogue 56 and paralogue 63 (Table 1) (Drouin et al. 1999). The latter result is expected, given that the null hypothesis inferred from analysis of the complete alignment suggests that paralogues 56 and 63 are recent duplicates of one another.

An ML-Distance Approximation

The use of ML to infer phylogenetic trees is computationally prohibitive for large data sets. We thus

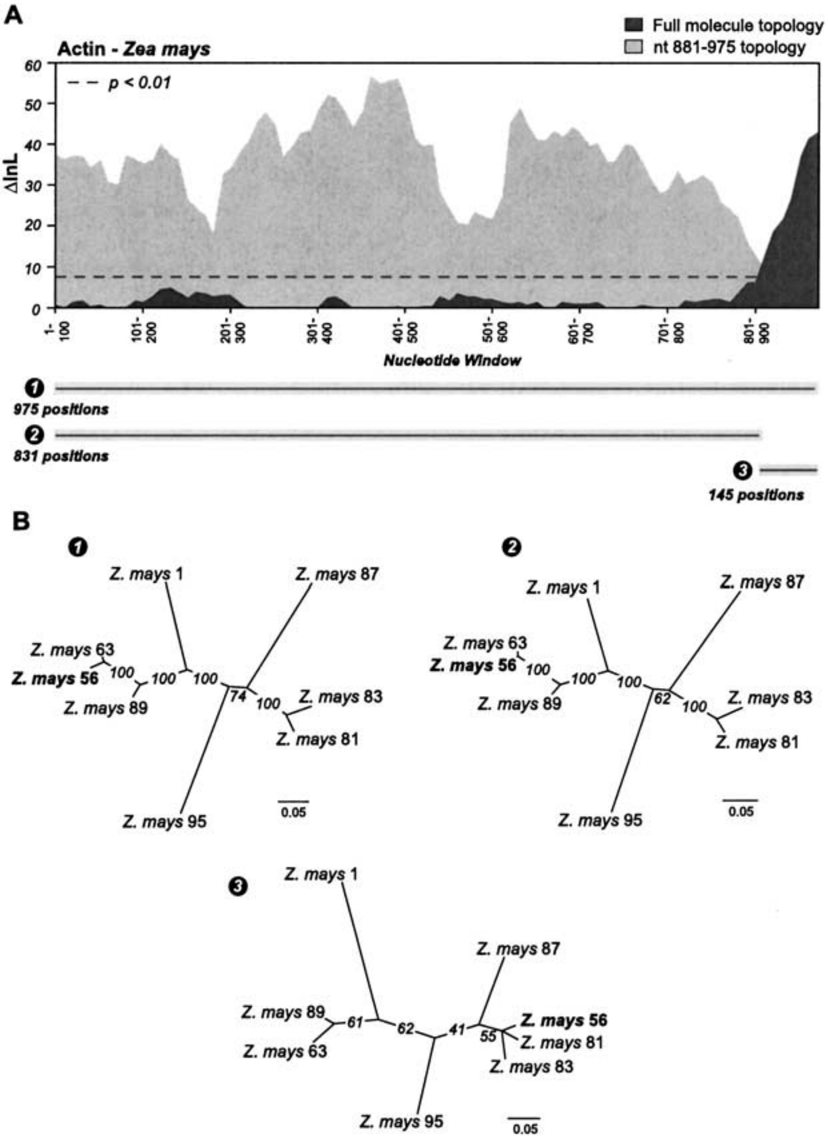

Fig. 6. Maximum likelihood (ML) sliding window analysis of *Zea mays* actin paralogues. **A** Δln*L* profiles inferred from an alignment of eight sequences and 975 nucleotide positions. The Δln*L* profile obtained when the topology inferred from the full molecule is used as the null hypothesis is plotted against that obtained when the tree inferred from the last 145 sites of the alignment was used. The *dashed line* indicates the estimate of the 0.99th quantile of the null distribution (Δln*L* = 7.68). Three discrete regions (data sets 1–3) of the alignment were selected for further phylogenetic analysis: their approximate positions and sizes are indicated. **B** ML trees inferred from the three data sets shown in **A**. ML bootstrap values are provided. Scale bars indicate the expected number of substitutions per site.

examined the efficacy of obtaining Δln*L* profiles from trees constructed with an ML-distance approximation. ML-distance Δln*L* profiles for the *Thermococcus*/*Pyrococcus* chaperonin, *Neisseria argF*, and maize actin data sets are shown in Fig. 7, plotted against those obtained from full ML analysis (Figs. 2, 5, and 6, respectively). On the whole, the results are very similar. A noticeable exception is the presence of the occasional negative Δln*L* value in the ML-distance profiles (e.g., the last window of the *Thermococcus*/*Pyrococcus* chaperonin data set; Fig. 7A). We examined these areas in isolation by phylogenetic analysis and determined that they correspond to regions of the alignment in which the ML and ML-distance trees conflict with one another (data not shown). As expected, the ML-distance method was much faster. For the *Neisseria argF* data set, the ML-distance approximation took 7 s on a 667-MHz Compaq EV6 (21160) Alpha processor, compared to 64 s for the full ML analysis. Null distributions inferred under the ML-distance approximation were similar to those obtained using ML (data not shown).

Discussion

We have presented an ML sliding window method that detects conflicting phylogenetic signals in nucleotide sequence alignments. A wealth of molecular data now suggests that intragenic recombination has been a major factor in the evolution of genes, and from the results presented here and elsewhere, it is clear that a well-resolved whole gene phylogeny cannot be taken as evidence that no recombination has occurred. Our method is meant to complement "standard" phylogenetic analyses by providing a statistically rigorous visual way of identifying anomalously evolving regions. As we have shown, subsequent analysis of such regions can provide insight into the nature of the

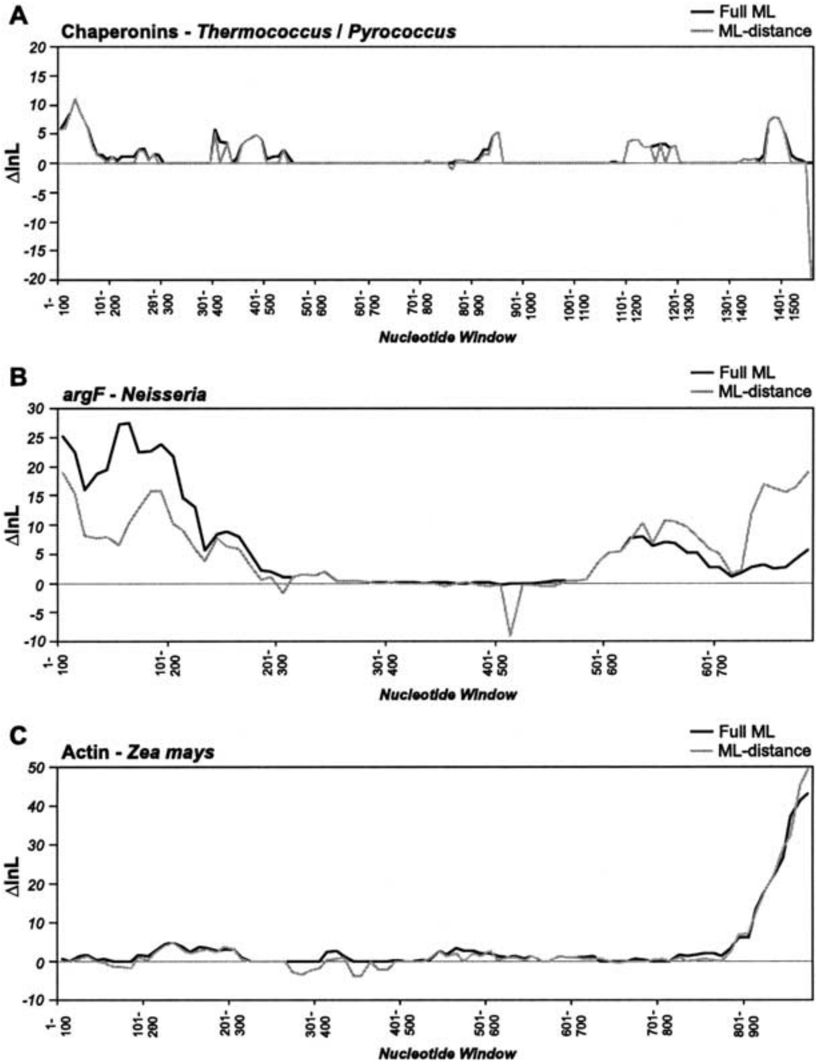

Fig. 7. Comparison of maximum likelihood (ML) and ML-distance sliding window analyses. **A**, **B**, **C** $\Delta\ln L$ profiles inferred from the chaperonin, *argF*, and actin data sets, respectively. $\Delta\ln L$ profiles obtained using the ML-distance approximation (*gray line*) are plotted against those obtained from the full ML analysis (*black line*).

evolutionary processes that have given rise to the anomaly. Using the "fixed-branch lengths" option, the method can also be used to detect gene conversions that do not result in changes in topology, as demonstrated for the methanogen chaperonin data set.

Our approach is similar to that of Grassly and Holmes (1997) in that we employ a sliding window method to identify regions of an alignment which do not fit with the ML topology inferred from the complete sequence. However, PLATO identifies regions that exhibit the lowest average likelihood relative to the rest of the molecule under a single null phylogeny. In contrast, we use a likelihood ratio test in which the difference between the log-likelihood of the ML tree for a given subset of the alignment and the log-likelihood of this region given a null phylogenetic hypothesis is used as a measure of conflict. Therefore, unlike PLATO, our method is relatively insensitive to regions of alignments with inherently low likelihoods, such as those evolving under positive selection or those that deviate significantly from the likelihood model used in the analysis. It is interesting that in three of the four euryarchaeal chaperonin data sets examined independently, PLATO—but not LIKEWIND—identified an area at the extreme 3′ end of the alignment as an anomalously evolving region (nt postions 1506–1554, 1509–1533, and 1528–1557 in the *Thermococcus*/*Pyrococcus*, methanogen, and *Thermoplasma* data sets, respectively; Table 1). The GTR rate matrix, P_{INV}, and Γ shape parameter estimates inferred from these regions did in fact differ significantly from those obtained from their respective complete alignments. As discussed above, the differences in the results obtained with LIKEWIND and PLATO are not due simply to the fact that LIKEWIND uses a slightly more complex model of sequence evolution.

Similarly, LIKEWIND identified a number of regions with demonstrably incongruent phylogenetic signals that PLATO did not detect. Two portions of the *Thermococcus*/*Pyrococcus* chaperonin alignment were flagged as anomalous by LIKEWIND but not PLATO, as were three regions of the methanogen chaperonin data set, one region in the halophile chaperonin data set, and one area in the *argF* data set (see Table 1). Although the reason for the failure of PLATO in these instances is not clear, in some cases it may be related to the program's reliance on finding discrete regions of an alignment which have a much lower average site likelihood within a given window than outside it. This procedure makes it inherently more difficult to identify anomalous regions when there is more than one such region. When multiple regions of "spatial phylogenetic variation" are present in an alignment, the average site likelihoods both inside and outside a window can be aberrantly low under the ML tree, and their ratio need not be particularly high. Indeed, in two of the cases where PLATO did not identify the phylogenetic discordant regions, there were multiple such regions (Table 1). In any case, a recent simulation study of recombination detection methods indicated that of the four methods tested (one of which was GENECONV), PLATO was the least powerful in detecting recombination over a wide range of recombination rates in a coalescent model (Brown CJ et al. 2001). For this reason, it is perhaps not surprising that PLATO failed to identify several clear cases of phylogenetic incongruence in the data sets analyzed in this study.

Our method clearly works best on alignments in which a dominant phylogenetic signal is present. However, the approach is not limited to the analysis of "well-resolved" phylogenies, since $\Delta\ln L$ profiles can be inferred under any *a priori* null hypothesis. If $\Delta\ln L$ profiles that are inferred under two null hypotheses possess peaks in the same region of an alignment, this suggests that a third topology is more consistent with the data in that region. This was found to be the case near the 3′ end of *Neisseria argF* alignment (see Figs. 5A and B). On the other hand, the converse situation in which $\Delta\ln L$ plots inferred under completely different null hypotheses have low $\Delta\ln L$ values in the same region suggests that the data in this portion of the alignment are phylogenetically uninformative. This could be due to either an extremely high or an extremely low degree of sequence conservation.

We found that LIKEWIND is largely unaffected by the presence of distantly related sequences in the alignment. Drouin et al. (1999) noted that GENECONV may perform poorly when one or more divergent sequences are present, as their presence masks the presence of silent polymorphic sites that would otherwise be selected for analysis. Indeed, GENECONV failed to detect recombination in the *Thermococcus*/*Pyrococcus* and halophile chaperonin data sets identified using LIKEWIND and/or PLATO, as well as in the *Neisseria argF* data set. This was also found to be the case in a recent analysis of chaperonin paralogues in crenarchaeotes. The method described here was used to identify gene conversions between paralogues sharing less than 50% sequence identity (Archibald and Roger 2002), conversions that were not detected using PLATO or the least-squares method (TOPAL) of McGuire and Wright (2000).

It is often desirable to be able to determine precisely the boundaries of anomalously evolving regions. In this respect, our approach is less precise than that of Grassly and Holmes (1997) or McGuire et al. (1997), the primary purpose of which is to detect putative recombination breakpoints. We used a sliding window size of 100 nucleotides for the LIKEWIND analyses, as this proved to be a reasonable balance between maximizing the ability to pinpoint anomalous regions and retaining sufficient phylogenetic signal for accurate results. Nevertheless, in cases where results were directly comparable, the boundaries predicted using the ML sliding window approach were often within 20 nucleotides of those predicted by GENECONV and PLATO (see Table 1).

At present, the method described in this paper is limited only by the large computational burden associated with full maximum likelihood. While we have performed ML sliding window analyses on data sets containing as many 21 sequences (data not shown), determining the significance of the resulting $\Delta\ln L$ profiles with a sufficient number of parametric bootstrap replicates is a daunting task. Such analyses often took between 24 and 48 h to complete on a 667-MHz Compaq EV6 (21160) Alpha processor, compared to seconds or minutes for an analysis of the same data set using PLATO or GENECONV. However, we found that inferring likelihoods from ML-distance trees produced $\Delta\ln L$ profiles comparable to those obtained with a full ML analysis, and as expected, these analyses were much faster. This should allow the method to be used in the analysis of much larger data sets, e.g., more than 40 sequences. At present, a data set-dependent bug in the latest release of PAUP* (b8) prevents us from investigating the ML-distance approximation in a rigorous fashion; once this problem is circumvented, the reliability and performance of the ML-distance option will be pursued further.

With respect to the phylogeny of chaperonins, the lineage-specific gene duplication pattern inferred for euryarchaeotes (Fig. 1) (Archibald et al. 1999, 2001) should, in light of the data herein, be seen as compatible with a scenario of somewhat fewer gene duplications combined with periodic complete gene conversion. Complete gene conversions would not be detected by any of the methods used here and would be indistinguishable from recent duplication. Nevertheless, the highest degree of amino acid identity shared between subunits encoded in the same archaeal genome is only 80.6% (*Thermococcus* sp. KS-1 *a* and *b*), suggesting that, if complete gene conversions between duplicate genes have occurred, they have not happened recently. It certainly appears to be the case that, as has been observed for numerous other gene families, small-scale intragenic recombination has played an important role in the evolution of the chaperonins.

The inability of many single-gene analyses to resolve important phylogenetic issues has led to the current trend of analyzing concatenated datasets (e.g., Baldauf et al. 2000; Brown JR et al. 2001). However, the lateral transfer of genes among genomes (Ochman et al. 2000) and underlying differences in the processes of molecular evolution from gene to gene (Huelsenbeck and Bull 1996) suggest that concatenation should be performed with caution. It is also now clear that intragenic recombination is a major concern: not only should some genes not be analyzed as part of a single data set, but discrete regions of individual genes may be incompatible with a single underlying evolutionary history. It is thus essential that the methods used to detect recombination or conflicting phylogenetic signals within and between genes and genomes continue to be improved.

Acknowledgments. We wish to thank E. Holmes for generously providing an alignment of *Neisseria argF* genes and G. Drouin for an alignment of Maize actin paralogues. We also thank A. Simpson for analytical suggestions. We thank J. Andersson and C. Nesbø for review of the manuscript, members of the Roger and Doolittle labs for helpful discussion, and two anonymous reviewers for their useful suggestions. We also thank E. Susko and C. Field for help with validating our statistical procedures. This work was supported by NSERC Grant 227085-00 awarded to A.J.R. J.M.A. was supported by a CIHR Doctoral Research Award and, subsequently, NSERC Grant 227085-00 awarded to A.J.R. A.J.R. thanks the CIAR Program in Evolutionary Biology for Fellowship support.

References

Archibald JM, Roger AJ (2002) Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J Mol Biol 316:1041–1050

Archibald JM, Logsdon JM, Doolittle WF (1999) Recurrent paralogy in the evolution of archaeal chaperonins. Curr Biol 9:1053–1056

Archibald JM, Blouin C, Doolittle WF (2001) Gene duplication and the evolution of group II chaperonins: implications for structure and function. J Struct Biol 135:157–169

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977

Bisercic M, Feutrier JY, Reeves PR (1991) Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: Evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J Bacteriol 173:3894–3900

Brown CJ, Garner EC, Dunker AK, Joyce P (2001) The power to detect recombination using the coalescent. Mol Biol Evol 18:1421–1424

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nature Genet 28:281–285

Bukau B, Horwich AL (1998) The Hsp70 and Hsp60 chaperone machines. Cell 92:351–366

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073

Drouin G, Prat F, Ell M, Clarke GD (1999) Detecting and characterizing gene conversions between multigene family members. Mol Biol Evol 16:1369–1390

DuBose RF, Dykhuizen DE, Hartl DL (1988) Genetic exchange among natural isolates of bacteria: Recombination within the *phoA* gene of *Escherichia coli*. Proc Natl Acad Sci USA 85:7036–7040

Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268

Feil E, Zhou J, Maynard Smith J, Spratt BG (1996) A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: Evidence for extensive interspecies recombination within *adk*. J Mol Evol 43:631–640

Fitch DHA, Goodman M (1991) Phylogenetic scanning: A computer assisted algorithm for mapping gene conversions and other recombinational events. CABIOS 7:207–215

Gangloff S, Zou H, Rothstein R (1996) Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. EMBO J 15:1715–1725

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239–247

Gutsche I, Essen LO, Baumeister W (1999) Group II chaperonins: New TRiC(k)s and turns of a protein folding machine. J Mol Biol 293:295–312

Halter R, Pohlner J, Meyer TF (1989) Mosaic-like organization of IgA protease genes in *Neisseria gonorrhoeae* generated by horizontal genetic exchange in vivo. EMBO J 8:2737–2744

Hein J (1993) A heuristic method to reconstruct the history of sequences subject to recombination. J Mol Evol 36:396–405

Huelsenbeck JP, Bull JJ (1996) A likelihood ratio test to detect conflicting phylogenetic signal. Syst Biol 45:92–98

Hughes AL (1995) Origin and evolution of HLA class I pseudogenes. Mol Biol Evol 12:247–258

Liao D (2000) Gene conversion drives within genic sequences: Concerted evolution of ribosomal RNA genes in bacteria and archaea. J Mol Evol 51:305–317

McGuire G, Wright F (2000) TOPAL 2.0: Improved detection of mosaic sequences within multiple alignments. Bioinformatics 16:130–134

McGuire G, Wright F, Prentice MJ (1997) A graphical method for detecting recombination in phylogenetic data sets. Mol Biol Evol 14:1125–1131

McGuire G, Wright F, Prentice MJ (2000) A Bayesian model for detecting past recombination events in DNA multiple alignments. J Comput Biol 7:159–170

Moniz de Sá M, Drouin G (1996) Phylogeny and substitution rates of angiosperm actin genes. Mol Biol Evol 13:1198–212

Nelson KE, Clayton RA, Gill SR, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H (2000) Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA 97:12176–12181

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Rambaut A, Grassly NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235–238

Ranson NA, White HE, Saibil HR (1998) Chaperonins. Biochem J 333:233–242

Rudikoff S, Fitch WM, Heller M (1992) Exon-specific gene correction (conversion) during short evolutionary periods: Homogenization in a two-gene family encoding the beta-chain constant region of the T-lymphocyte antigen receptor. Mol Biol Evol 9:14–26

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Scott AF, Heath P, Trusko S, Boyer SH, Prass W, Goodman M, Czelusniak J, Chang LY, Slightom JL (1984) The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. Mol Biol Evol 1:371–389

Smith JM, Dowson CG, Spratt BG (1991) Localized sex in bacteria. Nature 349:29–31

Smith NH, Holmes EC, Donovan GM, Carpenter GA, Spratt BG (1999) Networks and groups within the genus *Neisseria*: analysis of *argF*, *recA*, *rho*, and 16S rRNA sequences from human *Neisseria* species. Mol Biol Evol 16:773–783

Spratt BG (1988) Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae*. Nature 332:173–176

Spratt BG, Zhang QY, Jones DM, Hutchison A, Brannigan JA, Dowson CG (1989) Recruitment of a penicillin-binding protein gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. Proc Natl Acad Sci USA 86:8988–8992

Spratt BG, Bowler LD, Zhang QY, Zhou J, Smith JM (1992) Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J Mol Evol 34:115–125

Stephens JC (1985) Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Swofford DL (1998) PAUP*: Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, MA

Weiller GF (1998) Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol 15:326–335

Wiuf C, Christensen T, Hein J (2001) A simulation study of the reliability of recombination detection methods. Mol Biol Evol 18:1929–1939

Zhou J, Spratt BG (1992) Sequence diversity within the *argF*, *fop* and *recA* genes of natural isolates of *Neisseria meningitidis*: Interspecies recombination within the *argF* gene. Mol Microbiol 6:2135–2146