PhyloCSF: a comparative genomics method to distinguish protein-coding and non-coding regions

Michael F. Lin*‡ Irwin Jungreis* Manolis Kellis*†‡

Abstract

As high-throughput transcriptome sequencing provides evidence for novel transcripts in many species, there is a renewed need for accurate methods to classify small genomic regions as protein-coding or non-coding. We present PhyloCSF, a novel comparative genomics method that analyzes a multi-species nucleotide sequence alignment to determine whether it is likely to represent a conserved protein-coding region, based on a formal statistical comparison of phylogenetic codon models. We show that PhyloCSF's classification performance in 12-species *Drosophila* genome alignments exceeds all other methods we compared in a previous study, and we provide a software implementation for use by the community. We anticipate that this method will be widely applicable as the transcriptomes of many additional species, tissues, and subcellular compartments are sequenced, particularly in the context of ENCODE and modENCODE.

Preprint as of August 17, 2010

Introduction

High-throughput transcriptome sequencing is yielding precise structures for novel transcripts in many species, including mammals [1]. Accurate computational methods are needed to classify these transcripts and the corresponding genomic exons as likely to be protein-coding or non-coding, even if the transcript models are incomplete or if they only reveal novel exons of already-known genes. In addition to analyzing novel transcript models, such methods also have applications in evaluating and revising existing gene annotations [2, 3, 4, 5, 6], and as input features for *de novo* gene structure predictors [7, 8]. We have previously [9] compared numerous methods for determining whether an exon-length nucleotide sequence is likely to be protein-coding or non-coding, including single-sequence metrics based on the genome of interest only, and also comparative genomics metrics based on alignments with orthologous regions in the genomes of related species.

Among the comparative methods benchmarked in our previous study, one of our original contributions was Codon Substitution Frequencies (CSF), which assigns a score to each codon substitution observed in the input alignment based on the relative frequency of that substitution in known coding and non-coding regions. We showed that CSF is highly effective, performing competitively with a phylogenetic modeling approach with much less computational expense, and indeed we have applied it successfully in flies [10, 3], fungi [5], and mammals [4, 11, 1]. However, as discussed in our previous work, CSF has certain drawbacks arising from its *ad hoc* scheme for combining

*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. 32 Vassar St. 32-D510, Cambridge, MA 02139
†The Broad Institute. 7 Cambridge Center, Cambridge, MA 02142
‡Correspondence: {mikelin,manoli}@mit.edu

evidence from multiple species. For example, it makes only partial use of the evidence available in a multi-species alignment, and it produces a score lacking a precise theoretical interpretation, meaningful only relative to its empirical distributions in known coding and non-coding regions.

This preprint introduces a rigorous reformulation of CSF, which frames the evaluation of a given alignment as a statistical model selection problem, choosing between phylogenetic codon models estimated from known coding and non-coding regions as the best explanation for the alignment. This new "PhyloCSF" method fully leverages multiple alignments in a phylogenetic framework, produces meaningful likelihood ratios as its output, and rests upon the sweeping theoretical foundation for statistical model comparison. At the same time, it maintains certain advantages of CSF compared to existing phylogenetic methods. By reanalyzing the classification datasets from our original study, we show that PhyloCSF outperforms all of the other methods we previously benchmarked. Lastly, we briefly describe our software implementation, which we make publicly available.

Phylogenetic tests for distinguishing coding and non-coding regions

Our new PhyloCSF method builds upon the well-established theoretical framework for statistical phylogenetic model comparison. In this context, phylogenetic models are generative probabilistic models that produce alignments of molecular sequences, specifying a prior distribution over a common ancestral sequence, the topology and branch lengths of a phylogenetic tree relating the descendants, and a substitution process along each branch giving the rates (per unit branch length) at which each character changes to any other. In phylogenetic model comparison, we wish to choose between two competing models as the better explanation for a given alignment. A standard approach is to decide based on the *likelihood ratio* between the two models, which quantifies how much more probable the alignment is under one model than the other. This general approach has been used to explore many different aspects of the evolution of protein-coding genes, as recently reviewed in [12] and [13].

For distinguishing coding and non-coding regions, we design one phylogenetic model to represent the evolution of codons in protein-coding genes, and another to represent the evolution of nucleotide triplet sites in non-coding regions. These models may have one or more parameters θ that adjust them to the genomic region of interest, e.g. the background mutation rate or G+C content. To analyze a given alignment A of extant sequences, we first determine the probability of the alignment under the maximum likelihood estimate (MLE) of the parameters for the coding model, $p_C = \max_{\theta_C} \Pr(A|\,\text{Coding}, \theta_C)$. Then we do the same for the non-coding model to obtain $p_N = \max_{\theta_N} \Pr(A|\,\text{Noncoding}, \theta_N)$. Finally, we decide that the alignment is likely to represent a protein-coding region if the log-likelihood ratio $\Lambda = \log \frac{p_C}{p_N}$ is sufficiently high. The precise cutoff can be chosen to achieve a certain level of statistical significance, based on known asymptotic convergence properties of the log-likelihood ratio statistic [14, 15, 16], or it can be chosen empirically based on classification performance in a test set; we focus on the latter strategy in this work.

The d_N/d_S test

A standard method for detecting purifying selection on protein-coding sequences is to test for evidence that non-synonymous substitutions occur at a slower rate than synonymous substitutions. In the widely used PAML implementation of this test [17, 18], the codon frequencies π and the ratio of transition to transversion rates κ determine all triplet substitution rates in the background/non-coding model, while the coding model additionally supposes that non-synonymous codon substitution rates are reduced relative to synonymous rates by a scale factor ω (also called

d_N/d_S). PAML takes the phylogenetic tree topology as input, and estimates the branch lengths, π, κ, and ω for each alignment. The log-likelihood ratio between the coding and non-coding models can then be obtained from PAML's output. (By convention, the log-likelihood ratio is set to zero if the estimated $\omega \geq 1$.)

Our previous work [9] showed this to be one of the best comparative methods for distinguishing coding and non-coding regions, outperforming our CSF method according to standard classification error measures. Notably however, the d_N/d_S test performed worse than CSF for short regions ($\leq$ 180nt). This is not surprising since PAML was designed for evolutionary analysis of complete open reading frames, not short exon-length regions, which probably provide too little information to reliably estimate both the branch lengths and codon frequencies in addition to the two rate parameters.

PhyloCSF

Our new PhyloCSF method differs from the standard d_N/d_S test in two main ways. First, it takes advantage of recent advances in phylogenetic codon modeling methods that enable much more detailed representations of coding and non-coding sequence evolution. Specifically, while the d_N/d_S test uses only a few parameters to model the rates of all possible codon substitutions [17], PhyloCSF uses two *empirical codon models* (ECMs) based on independent parameters for essentially all such rates [19], one estimated from alignments of many known coding regions, and the other from non-coding regions. By comparing these two rich evolutionary models, PhyloCSF can observe many additional informative features of a given alignment compared to the d_N/d_S test. For example, the coding ECM captures not only the decreased overall rate of non-synonymous substitutions, but also the different rates of specific non-synonymous substitutions reflecting the chemical properties of the amino acids. (Earlier codon modeling approaches also incorporate amino acid distances, e.g. [20], but to our knowledge, these are not widely used for discriminating between coding and non-coding regions.) Also, our ECMs explicitly model the very different rates of substitutions giving rise to stop codons in coding and non-coding regions.

Second, PhyloCSF also relies on genome-wide training data to provide prior information about the branch lengths in the phylogenetic tree and the codon frequencies, rather than attempting to re-estimate these *a priori* even in very short alignments. PhyloCSF assumes a fixed tree "shape" based on the genome-wide MLEs of the branch lengths, and estimates only two scale factors ρ_C and ρ_N, applied uniformly to all of the branch lengths in the coding and non-coding models respectively, for each individual region analyzed. This allows the method to accomodate some region-specific rate variation and reduces its sensitivity to the absolute degree of conservation, without greatly increasing the parameterization at the expense of reliability for short regions.

In summary, PhyloCSF relies on two ECMs fit to genome-wide training data (Figure 1A), which include estimates for the branch lengths, codon frequencies, and codon substitution rates for known coding and non-coding regions. To evaluate a given nucleotide sequence alignment (Figure 1B,C), PhyloCSF (1) determines the MLE of the scale factor ρ on the branch lengths for each of these models, (2) computes the likelihood of each model (the probability of the alignment under the model) using the MLE of the scale factor, and (3) reports the log-likelihood ratio between the coding and non-coding models.

PhyloCSF outperforms other methods

To evaluate the discriminatory power of our new method, we applied it to the same datasets used in our previous study [9], and also benchmarked its performance in the same way. Briefly, the

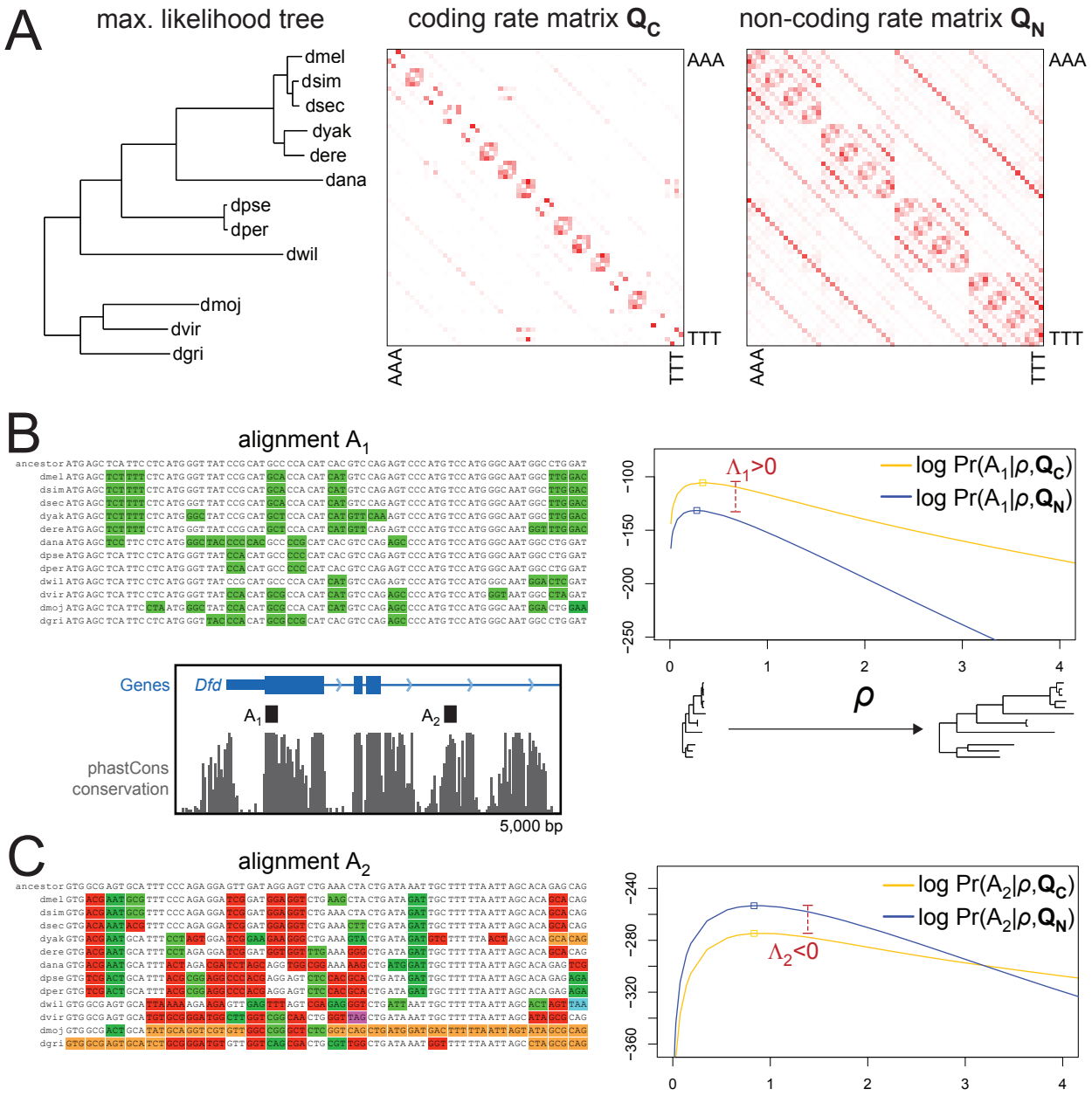

Figure 1: PhyloCSF method overview. **(A)** PhyloCSF uses phylogenetic codon models estimated from genome-wide training data based on known coding and non-coding regions. These models include a phylogenetic tree and codon substitution rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ for coding and non-coding regions, respectively, shown here for 12 *Drosophila* species. $\mathbf{Q_C}$ captures the characteristic evolutionary signatures of codon substitutions in conserved coding regions, while $\mathbf{Q_N}$ captures the typical evolutionary rates of triplet sites in non-coding regions. **(B)** PhyloCSF applied to a short region from the first exon of the *D. melanogaster* homeobox gene *Dfd*. The alignment of this region shows only synonymous substitutions compared to the inferred ancestral sequence (green). Using the maximum likelihood estimate of a scale factor ρ applied to the assumed branch lengths, the alignment has higher probability under the coding model than the non-coding model, resulting in a positive log-likelihood ratio Λ. **(C)** PhyloCSF applied to a conserved region within a *Dfd* intron. In contrast to the exonic alignment, this region shows many non-synonymous substitutions (red), nonsense substitutions (blue, purple), and frameshifts (orange). The alignment has lower probability under the coding model, resulting in a negative score.

datasets consist of known protein-coding regions and random non-coding regions (about 50,000 total regions) in the genome of the fruitfly *Drosophila melanogaster*, aligned with eleven other *Drosophila* species using MULTIZ [21, 22, 10]. The lengths of the regions in both the coding and non-coding sets match the length distribution of fly coding exons. Consistent with our previous work, we trained and applied PhyloCSF on this dataset using four-fold cross-validation, to ensure that any observed performance differences are not due to overfitting. We assessed the results by examining ROC curves and computing the minimum average error (MAE), the average false positive and false negative rates at the cutoff that minimizes this average. To compare the power of the methods for short exons specifically, we additionally computed these benchmarks only for the 37% of examples from 30 to 180 nucleotides in length.

These benchmarks showed that PhyloCSF outperforms the other comparative methods we previously benchmarked, essentially dominating them at good sensitivity/specificity tradeoffs (Figure 2A). PhyloCSF's overall MAE was 19% lower than that of the Reading Frame Conservation metric, 15% lower than our older CSF method, and 8% lower than the d_N/d_S test. PhyloCSF also clearly outperformed the other methods for short exons (Figure 2B), with an MAE 11% lower than the d_N/d_S test. These results show that PhyloCSF provides superior power to distinguish coding and non-coding regions based on multi-species genome alignments.

Implementation and availability

To facilitate the use of PhyloCSF by the community, we provide an implementation that evaluates an input sequence alignment in Multi-FASTA format and reports the resulting log-likelihood ratio in units of decibans. The program supports batch processing of many alignments, and can either evaluate each alignment as-is or search for high-scoring open reading frames within. We also provide the ECMs and other parameter settings for several phylogenies, including mammals and flies. The Objective Caml source code and executables for popular platforms are available at: `http://compbio.mit.edu/PhyloCSF`

Discussion

We have introduced PhyloCSF, a comparative genomics method for distinguishing protein-coding and non-coding regions, and shown that it outperforms other methods. In addition to its superior discriminatory power, PhyloCSF is far more theoretically attractive than our older CSF method and other *ad hoc* metrics, relying on a formal statistical comparison of phylogenetic codon models. On the other hand, PhyloCSF and CSF produce highly correlated scores (Pearson coefficient 0.95 in our dataset), and the new method is much more computationally demanding. It should also be noted that CSF and PhyloCSF, unlike many of the other methods compared in our previous study, require extensive genome-wide training data from known coding and non-coding regions, which can present an obstacle in genomes outside of well-studied phylogenies such as mammals and flies.

The classification approach described here, in which we are given individual sequence alignments and must decide whether each one represents a coding or non-coding region, is complementary to *de novo* comparative gene predictors that attempt to parse the complete primary genome sequence into exon-intron structures and intergenic regions. In particular, we suggest that the classification approach is more naturally applicable to transcript models reconstructed from expression evidence such as mRNA-Seq, where the exon-intron structure is essentially known. Moreover, the rapidly decreasing cost of such transcriptome sequencing is arguably reducing the need for *de novo* gene predictors, although they will remain useful for predicting possible lowly- or rarely-expressed genes.

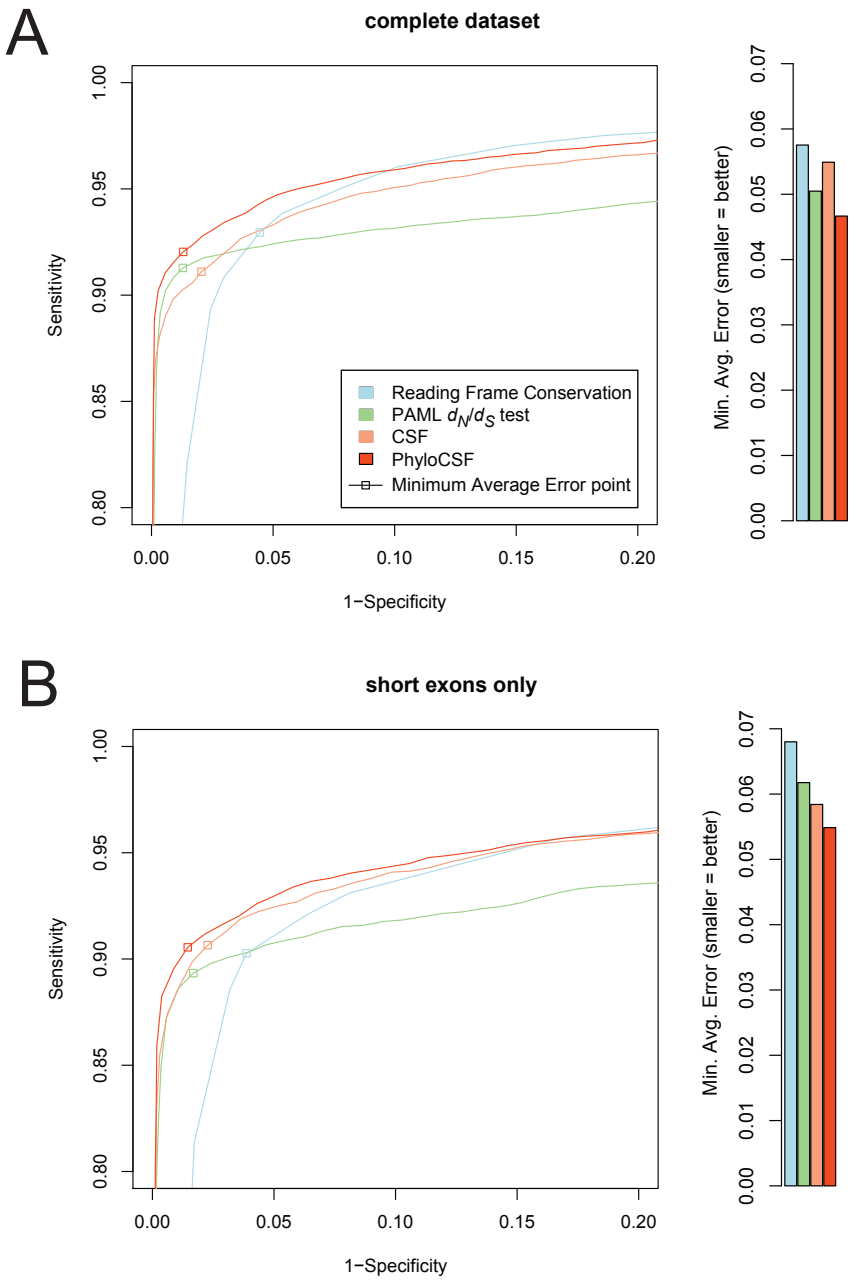

Figure 2: PhyloCSF performance benchmarks. **(A)** ROC curves and error measures for distinguishing coding and non-coding regions in a dataset of approximately 50,000 regions from *D. melanogaster* aligned to 11 other fly genomes. PhyloCSF clearly outperforms the other methods. **(B)** As in (A), but only for the 37% of regions in the dataset between 30 and 180 nucleotides in length.

Methods

PhyloCSF

PhyloCSF's trained parameters include a phylogenetic tree (with branch lengths) and two 64-by-64 codon rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ representing coding and non-coding sequence evolution, respectively, as reversible, homogeneous, continuous-time Markov processes. To evaluate a given alignment, we first evaluate the likelihood of the coding model as follows. First, we define an alignment-specific parameter ρ_C that operates as a scale factor applied to all of the branch lengths in the predefined tree. Given a setting of ρ_C, the substitution probability matrix along any branch with length t is given by $\mathbf{P} = \exp(t\rho_C\mathbf{Q_C})$, and the probability of the full alignment can be efficiently computed using Felsenstein's algorithm [23], assuming independence of the codon sites, using the equilibrium frequencies implicit in $\mathbf{Q_C}$ as the prior distribution over the common ancestral sequence, and marginalizing out any gapped or ambiguous codons. We numerically maximize this probability over ρ_C to obtain the likelihood of the coding model p_C. We then evaluate the likelihood of the non-coding model p_N in the same way, using $\mathbf{Q_N}$ and an independent scale factor ρ_N, and report the log-likelihood ratio $\log \frac{p_C}{p_N}$ as the result.

Estimation of empirical codon models

To estimate the phylogenetic tree and the empirical rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ for the species of interest, we rely on sequence alignments of many known coding and random non-coding regions. Given this genome-wide training data, we optimize the parameters for the coding and non-coding models using an expectation-maximization approach. The E-step is carried out as previously described [24, 25]. In each M-step, we update the ECM exchangeability parameters using a spectral approximation method [26] and the branch lengths by gradient ascent on the expected log-likelihood function [25]. Meanwhile, the codon/triplet frequencies are fixed to their empirical averages in the training examples, and we assume a fixed species tree topology.

Acknowledgments

The authors thank Matthew D. Rasmussen and Manuel Garber for helpful comments and discussions. Funding for this work was provided by the National Institutes of Health (U54 HG004555-01) and the National Science Foundation (DBI 0644282).

References

[1] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. Nat Biotech 28: 503-510.

[2] Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-254.

[3] Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, et al. (2007) Revisiting the protein-coding gene catalog of drosophila melanogaster using 12 fly genomes. Genome research 17: 1823-1836.

[4] Clamp M, Fry B, Kamal M, Xie X, Cuff J, et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences 104: 19428-19433.

[5] Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, et al. (2009) Evolution of pathogenicity and sexual reproduction in eight candida genomes. Nature 459: 657-662.

[6] Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, et al. (2009) The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research 19: 1316-1323.

[7] Brent MR (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nature reviews Genetics 9: 62-73.

[8] Alioto T, Guig R (2008). State of the art in eukaryotic gene prediction.

[9] Lin MF, Deoras AN, Rasmussen MD, Kellis M (2008) Performance and scalability of discriminative metrics for comparative gene identification in 12 drosophila genomes. PLoS Comput Biol 4: e1000067.

[10] Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. Nature 450: 219-232.

[11] Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. Nature 458: 223-227.

[12] Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Molecular biology and evolution 26: 255-271.

[13] Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Briefings in bioinformatics 10: 97-109.

[14] Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 16: 1292-1299.

[15] Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Molecular biology and evolution 17: 798-803.

[16] Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57: 307-333.

[17] Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of Molecular Evolution 46: 409-418.

[18] Yang Z (2007) Paml 4: Phylogenetic analysis by maximum likelihood. Molecular biology and evolution 24: 1586-1591.

[19] Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Molecular biology and evolution 24: 1464-1479.

[20] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding dna sequences. Molecular biology and evolution 11: 725-736.

[21] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome research 14: 708-715.

[22] Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the drosophila phylogeny. Nature 450: 203-218.

[23] Felsenstein J (2004) Inferring Phylogenies. Sunderland, MA: Sinauer Associates, Inc.

[24] Holmes I, Rubin GM (2002) An expectation maximization algorithm for training hidden substitution models. Journal of Molecular Biology 317: 753-764.

[25] Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21: 468-488.

[26] Arvestad L, Bruno WJ (1997) Estimation of reversible substitution matrices from multiple pairs of sequences. Journal of Molecular Evolution 45: 696-703.