Simultaneous Gene Finding in Multiple Genomes

Stefanie König[1], Lars Romoth[1], Lizzy Gerischer[1], and Mario Stanke[1]

[1]**Institute for Mathematics and Computer Science, University of Greifswald, {stefanie.koenig,lars.romoth,lizzy.gerischer,mario.stanke}@uni-greifswald.de**

ABSTRACT

As whole genome sequencing is taking on ever-increasing dimensions, the new challenge is the accurate and consistent annotation of entire clades of genomes. We address this problem with a new approach to comparative gene finding that takes a multiple genome alignment of closely related species and simultaneously predicts the location and structure of protein-coding genes in all input genomes, thereby exploiting negative selection and sequence conservation. The model prefers potential gene structures in the different genomes that are in agreement with each other, or – if not – where the exon gains and losses are plausible given the species tree. We formulate the multi-species gene finding problem as a binary labeling problem on a graph. The resulting optimization problem is NP hard, but can be efficiently approximated using a subgradient-based dual decomposition approach. The proposed method was tested on a whole-genome alignment of 12 *Drosophila* species and its accuracy evaluated on *D. melanogaster*. The method is being implemented as an extension to the gene finder AUGUSTUS.

Keywords: gene finding, comparative genomics, genome annotation, dual decomposition

INTRODUCTION

With recent technologies in whole-genome sequencing, the sequencing of entire clades of genomes is in progress. For example, the Genome 10K Project launched in 2009 has taken on the task of sequencing the genomes of 10,000 vertebrate species (Genome 10K Community of Scientists, 2009). Other examples include the 5,000 Insect Genome Project (i5k) (Robinson et al., 2011) and the 1,000 Fungal Genomes Project of the JGI. The JGI has further more than 50 strains of *Brachypodium* (model grass) and is going to obtain genomes for about 50 switchgrass strains and more than 20 *Brassicaceae* (David Goodstein, personal communication).

The annotation of genomes, in turn, is a rather slow process. An important step is the identification of protein-coding genes. Although many automatic tools for gene finding are available, none of them is able to predict genes genome-wide without a substantial rate of wrong gene structures or missing genes. For instance, a survey from 2013 (Steijger et al., 2013) suggests that even the most accurate tools are merely predicting 48.53% of the genes (at least one splice form) in *Drosophila melanogaster* correctly, when using only RNA-Seq data as evidence. For a recent review on the subject, see Hoff and Stanke (2015).

Another evidence source besides transcriptome sequence is homology. One class of methods that exploit homology uses *previously identified* protein sequences from related species or from a database and performs a spliced alignment against a target genome. Examples are the ENSEMBL pipeline that uses amongst other tools GENEWISE Birney et al. (2004) for protein sequence-based gene prediction, and AUGUSTUS-PPX Keller et al. (2011). These approaches depend on the correctness of the input proteins, their similarity to the target clade and the overlap of the respective proteomes. They are usually suited only as one component of a whole-genome annotation pipeline. A second class of methods that exploit homology are comparative gene finders. These methods take two or more genome sequences as input and exploit that homolog genes have often a very similar gene structure. By aligning the genomes of related species, conserved regions become visible that are enriched in protein-coding exons but also other functional DNA.

Initial comparative approaches to gene finding include pair hidden Markov models (Pair HMMs) to simultaneously predict genes in exactly two input genomes (e.g. of human and mouse) (Meyer and Durbin, 2002; Alexandersson et al., 2003). However, the generalization of Pair HMMs to multiple genomes does not scale well and Pair HMMs appear to play no substantial role in current genome annotation.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1296v1 | CC-BY 4.0 Open Access | rec: 11 Aug 2015, publ: 11 Aug 2015

To take advantage of a multiple genome alignment, alternative approaches restrict gene finding to a *single target genome* and use an alignment between the target and multiple related genomes to inform gene finding in the target. Examples are CONTRAST (Gross et al., 2007) and N-SCAN (Gross and Brent, 2006). In particular, CONTRAST achieved striking results (58.6% sensitivity and 35.5% specificity for human on gene level). Despite the very good performance of comparative gene finding and the potential to combine homology evidence with evidence from transcriptome sequencing, CONTRAST and N-SCAN are rarely used for whole-genome annotation. Reasons may include the fact that both require an elaborate parameter training specific to the set of 'informant' genomes, that has to be repeated for every genome in the clade that should be annotated. A methodical disadvantage is further the restriction of gene finding to a single target genome. This has the drawback, that likely gene structures in the informant genomes are not taken into consideration when choosing a gene structure in the target genome.

We present a novel approach to comparative gene finding that simultaneously identifies genes in $k \geq 2$ genomes and that is suitable for the annotation of entire clades of genomes, e.g. the runtime is linear in the number of genomes k. We introduce a graph-theoretical framework and formulate the problem as a binary labeling problem on a graph. In general, exact inference in this model is not tractable, however, we can take advantage of the special structure of the graph that allows decomposition into two tractable sub problems: Finding longest paths in directed acyclic graphs (DAGs), and maximum *a-posteriori* probability (MAP) inference on trees. A subgradient-based *dual decomposition* approach is derived for approximate inference, guaranteeing an upper bound on the approximation error. Dual decomposition and more generally Lagrangian relaxation has already been applied to a variety of inference problems, e.g. for the multiple sequence alignment problem (Althaus and Canzar, 2008), *de novo* peptide sequencing (Andreotti et al., 2012), computer vision (Komodakis et al., 2011) and natural language processing (Rush et al., 2010).

The proposed method is implemented as an extension to the gene finder AUGUSTUS (Stanke et al., 2008) and in the following referred to as AUGUSTUS$_{\text{CGP}}$. The required inputs are the genomes of two or more species as well as an alignment and a phylogenetic tree of the genomes. With no further information, AUGUSTUS$_{\text{CGP}}$ infers gene structures *de novo* by only making use of the raw genomes and alignment information. AUGUSTUS$_{\text{CGP}}$ incorporates evidence for negative selection by computing an estimate for the ratio of nonsynonymous and synonymous substitutions $\omega = dN/dS$ for all considered candidate coding exons. Furthermore, AUGUSTUS$_{\text{CGP}}$ can incorporate additional evidence, e.g. from RNA-Seq and existing annotations. The latter is used for the special application of transferring a trusted annotation from a known genome to newly sequenced genomes. The performance of AUGUSTUS$_{\text{CGP}}$ for all three tasks - *de novo*/evidence-based gene finding and cross-species annotation transfer - is evaluated on 12 *Drosophila* genomes and discussed in the results section.

Training the parameters of AUGUSTUS$_{\text{CGP}}$ is not more expensive than hitherto for a single genome. The species-specific parameters are only learned for one representative in the clade (e.g. human in a mammalian clade) with no need for retraining when more genomes are added to the clade or removed. Apart from the species-specific parameters there are only few extra cross-species parameters to adjust such as rates for exon gain and loss.

METHODS

Here, we formally introduce the problem of comparative gene finding using a graph and a scoring function for all possible *joint gene structures* in k homologous sequences. The problem is NP-hard. Therefore, we derive an approximative algorithm based on dual decomposition for determining a joint gene structure with maximal score. Given the page constraint, we omit many details of the program and focus on the algorithmic part here. A separate publication describing the details of the scores is planned.

The Model of a Joint Gene Structure

Let us first consider a single genomic sequence g. The space of all possible gene structures [1] x in g can be modeled as paths from a source s to a sink ℓ in a weighted directed acyclic graph, which in the following is referred to as *gene structure graph*. For a conceptual example, see the graph for sequence 1 in Fig. 1. Nodes in the gene structure graph denote candidate exons. Directed edges represent candidate introns or intergenic regions and connect two nodes if they constitute a biologically meaningful sequence of exons. Both candidate exons and introns are obtained within

[1] We do not consider gene structures with overlapping transcripts, such as from alternative splicing. A gene structure may cover one or several genes, or even just intergenic region.

AUGUSTUS by random sampling of gene structures from the posterior distribution defined by a semi-Markov conditional random field. The sampling of gene structures in AUGUSTUS has previously been introduced to identify alternative transcripts (Stanke et al., 2006). In general, sampling yields just the most likely splicing variants, which do not sufficiently represent the space of all possible gene structures. To account for this, two adjustments are made. First, the posterior distribution is heated by raising its posterior probabilities to the power of $r \in (0,1]$ and subsequent renormalizing, $P_r(x) \propto (P(x))^r$. As a consequence, the sampling of less likely gene structures increases for $r < 1$, the sample of candidate exons is more inclusive and for candidate exons, that are frequently sampled, their heated posterior probability $P_r(x)$ is a more conservative estimate of the probability of being correct than in the original distribution P.

In addition, candidate exons are inserted into the graph that were not sampled. These are determined by all possible combinations of exon boundary signals (translation start/stop and donor/acceptor splice sites) that are within a given distance and that do not contain in-frame stop codons. The number of such candidate exons is within the same order of magnitude as the length of sequence g. To reduce run time and memory usage, candidate exons may be filtered by imposing a threshold on splice site scores.

The score of each source-sink path is a sum of node and edge weights. Both node and edge weights are real-valued functions of the posterior probability of the exon or intron as estimated by the relative sampling frequency of the corresponding candidate exons and introns, respectively. Exon candidates that are not sampled, are scored as if they have posterior probability 0. Furthermore, if extrinsic evidence is given, such as from RNA-Seq, then the weights indirectly depend on the evidence as candidate exons and introns that are supported by evidence typically achieve high posterior probabilities. The problem of finding an optimal gene structure in such a single genomic sequence, can be solved efficiently with standard algorithms for longest-paths problems.

Now, let us consider a syntenic region[2] consisting of k homologous sequences. Let $G^i(V^i, E^i)$ be the gene structure graph of sequence g_i $(i \in \{1,..,k\})$ with node set V^i and edge set E^i. The gene structure graphs are now combined into a single graph by connecting homologous candidate exons via phylogenetic trees as follows: Let $\sim$ denote an equivalence relation on $V = \cup_{i=1}^{k} V^i$, such that for $u \in V^i, v \in V^j, i \neq j, u \sim v$ if and only if both start and end positions of candidate exons u and v map to the same positions in the alignment. The relation $\sim$ partitions V into a set of equivalent classes, each of which is referred to as a *homologous exon candidate tuple* (HECT). All elements in a HECT are candidate exons that are putative homologs, meaning that they are believed to be derived from a common ancestor. The elements of singletons are candidate exons with no homologs in the other sequences. All exons in a HECT are linked by a phylogenetic tree by merging them with their counterparts (e.g. leaf nodes) in the tree. The tree is a copy of the input species tree in which the leaf node of species i is pruned if the HECT does not contain an exon candidate of species i.

Let $G(V \cup A, E_D \cup E_U)$ denote the *joint gene structure graph*, in which $V = \cup_{i=1}^{k} V^i$, A is the set of all ancestral exons (interior nodes of the phylogenetic trees), $E_D = \cup_{i=1}^{k} E^i$ is the set of all directed, 'intron' edges and E_U is the set of all undirected, phylogenetic edges in G. The joint gene structure graph comprises all possible gene structures of all k sequences (see Fig. 1). Loosely speaking, the aim is to choose exactly one gene structure, or equivalently source-sink path $s^i \rightsquigarrow \ell^i$, for each sequence $g_i, i = 1,..,k$. In the following, such a collection of k paths is also called a *joint gene structure*. In mathematical terms a joint gene structure is an assignment $\mathbf{x} = (x_1,...,x_n) \in \chi \subset \{0,1\}^n, n = |V| + |A|$ of all nodes in G. Observe that this formal definition of a joint gene structure includes the choice of ancestral exons. A node v is assigned to 1 if it is part of the joint gene structure and 0 otherwise. We will also say that v is *active* if $x_v = 1$ and *inactive* if $x_v = 0$. Likewise, an edge $(u,v) \in E_D$ is active if both u and v are active and there is no path from u to v that passes through active nodes other than u and v. The subset χ is the set of all assignments that obey the path property, e.g. each source node has exactly one outgoing active edge, each sink node has exactly one incoming active edge and all other nodes have an equal number (0 or 1) of incoming active edges and outgoing active edges.

The score $S(\mathbf{x})$ of a joint gene structure $\mathbf{x}$ has two components, a *horizontal*, species-specific score $h(\mathbf{x})$ and a *vertical*, cross-species score $v(\mathbf{x})$:

$$S(\mathbf{x}) = h(\mathbf{x}) + v(\mathbf{x}) \tag{1}$$

The horizontal score is the sum over all weights of active nodes $v \in V$ and active edges $e \in E_D$. The

[2] In general, there will be many, sometimes overlapping syntenic regions between the genomes or a subset of the genomes, each of which is an instance of the proposed method. These regions of synteny are determined within AUGUSTUS$_{\text{CGP}}$ by merging compatible alignment blocks in the input alignment to larger blocks of synteny. Note, however, that this is not a trivial task in itself, especially in cases where the genome assemblies are highly fragmented. Ideally, the syntenic regions are large enough to contain one ore more genes.

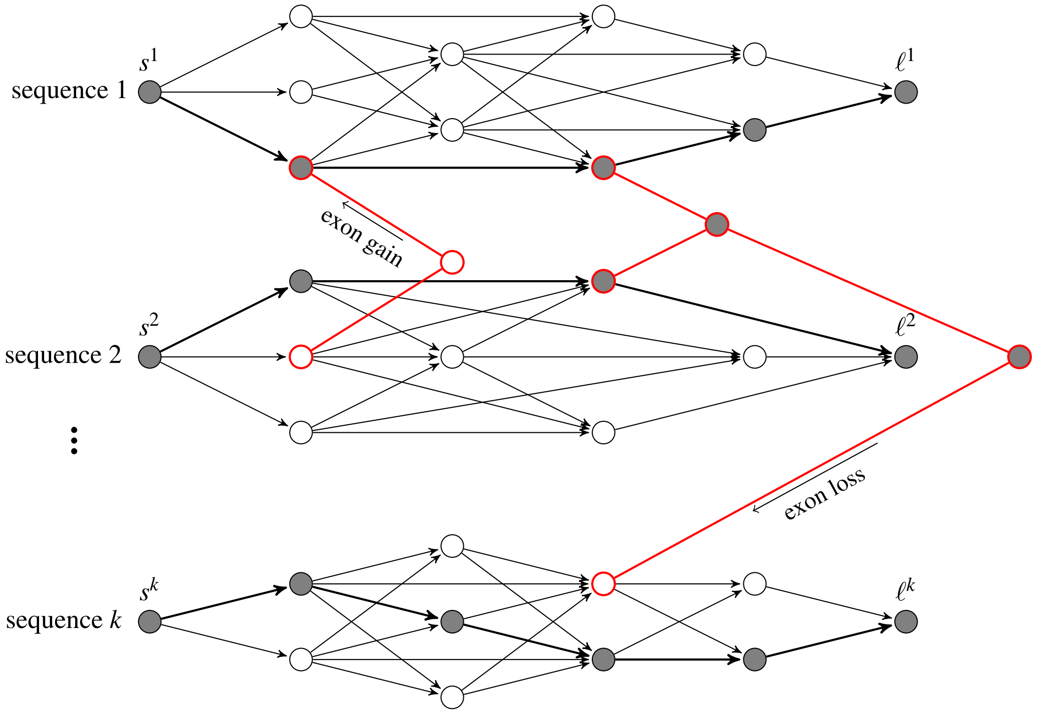

Figure 1. The joint gene structure graph G for a set of k homologous sequences. Nodes represent candidate exons. Directed edges represent candidate introns or intergenic regions. Each path from the source s^i to the sink ℓ^i is a possible gene structure in sequence i. Homologous candidate exons are at the same time leaf nodes of phylogenetic trees (red edges and nodes). A joint gene structure – a collection of k paths $s^i \rightsquigarrow \ell^i, i = 1,..,k$ – is sought or, equivalently, the corresponding binary labeling (●,○) of all nodes in G.

node and edge weights are the ones from gene finding in a single genomic sequence as described above. The vertical score is a sum over the trees in the graph, a function of the labels of all nodes in HECTs and can be split into a *feature score* and an *evolutionary score*. The feature score is a linear combination of different features of homologous candidate exons including selective pressure (estimated by $\omega = {}^{dN}\!/_{dS}$), phylogenetic diversity (sum of branch lengths in the tree that connects a HECT) and conservation (average Shannon entropy across all alignment columns in a HECT). It rewards candidate exons that show signs of negative selection ($\omega \ll 1$) and are conserved even across the more distant species. The coefficients of the linear combination are calculated using logistic regression and the R programming language (R Core Team, 2013). The evolutionary score is based on a continuous-time Markov process and assesses the evolutionary history of a joint gene structure in terms of exon gain and loss events along branches in the tree (see Fig 1). Similar models have been used previously for intron evolution (Csűrös, 2006). In this model, the gain or loss of an exon is generally expensive and penalized depending on the branch length and two rates for exon gain $\lambda \in \mathbb{R}_{>0}$ and exon loss $\mu \in \mathbb{R}_{>0}$. As a consequence, all candidate exons in a HECT are encouraged to agree on one assignment. If not, assignments are preferred that can be explained with few exon gain or loss events that are rather along long branches than along short branches.

Dual Decomposition

Finding an optimal joint gene structure $\mathbf{x}^*$ that maximizes the scoring function in (1)

$$S(\mathbf{x}^*) = \max_{\mathbf{x}} S(\mathbf{x}) \tag{2}$$

is an NP-complete problem, even if the vertical score is assumed to be a simple parsimony score penalizing exon gain and loss only. This can be shown by a reduction from the 3-colorability problem. For this reason, an approximative approach, known as *dual decomposition*, has been adopted, that makes use of the observation that the problem in (2) is decomposable into two easy sub problems:

$$\max_{\mathbf{y}_h,\mathbf{z}} h(\mathbf{y}_h,\mathbf{z}) \quad (3) \qquad \text{and} \qquad \max_{\mathbf{y}_v,\mathbf{z}} v(\mathbf{y}_v,\mathbf{z}) \quad (4)$$

Here, the assignment $\mathbf{x}$ is partitioned into three disjoint assignments $\mathbf{x} = (\mathbf{y}_h, \mathbf{y}_v, \mathbf{z})$, where $\mathbf{y}_h$ is an assignment of all candidate exons $\{v \in V \mid \nexists a \in A : \{v,a\} \in E_U\}$ with no homologs in the other species, $\mathbf{y}_v$ is an assignment of all ancestral exons $a \in A$, and $\mathbf{z}$ is an assignment of all homologous candidate exons $\{v \in V \mid \exists a \in A : \{v,a\} \in E_U\}$. Problem (3) maximizes over the horizontal score and is equivalent to finding an optimal gene structure in each of the k sequences individually. It can be solved efficiently with an algorithm for longest-path problems. Problem (4) maximizes over the vertical score and is equivalent to finding an optimal assignment of nodes in a set of disjoint trees. This can also be solved efficiently, for example with a variant of Felsenstein's pruning algorithm (Felsenstein, 2003). However, maximizing over the sum of the horizontal and vertical score (e.g., problem (2)) is hard, due to the *complicating variables* $\mathbf{z}$ that couple the two subproblems.

The Lagrangian Dual Problem

An equivalent formulation of problem (2), in which each subproblem has its own copy of complicating variables, is

$$\max_{\mathbf{y}_h,\mathbf{y}_v,\mathbf{z}_h,\mathbf{z}_v} h(\mathbf{y}_h,\mathbf{z}_h) + v(\mathbf{y}_v,\mathbf{z}_v), \text{ s.t. } \mathbf{z}_h = \mathbf{z}_v \tag{5}$$

The constraint $\mathbf{z}_h = \mathbf{z}_v$ ensures that the two sub problems agree on their copies of complicating variables. In the next step, Langrangian relaxation is applied by dropping the constraint and moving it into the objective function

$$L(\boldsymbol{\lambda}) = \max_{\mathbf{y}_h,\mathbf{z}_h} \left(h(\mathbf{y}_h,\mathbf{z}_h) + \boldsymbol{\lambda}^\intercal \mathbf{z}_h\right) + \max_{\mathbf{y}_v,\mathbf{z}_v} \left(v(\mathbf{y}_v,\mathbf{z}_v) - \boldsymbol{\lambda}^\intercal \mathbf{z}_v\right) \tag{6}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathbf{z}|}$ is the set of Lagrange Multipliers, which can be regarded as penalty for violating the constraint $\mathbf{z}_h = \mathbf{z}_v$, and $L(\boldsymbol{\lambda})$ is the Lagrangian Dual function. Since the Lagrangian Dual is an upper bound on $S(\mathbf{x}^*)$ for any $\boldsymbol{\lambda}$, the tightest upper bound, e.g. the set of Lagrange Multipliers $\boldsymbol{\lambda}^*$ that minimizes the Lagrangian Dual function,

$$S(\mathbf{x}^*) \leq L(\boldsymbol{\lambda}^*) = \min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda})$$

is sought. This is also known as the dual problem. Note that the Lagrangian Dual function is convex but, in general, not differentiable. Thus, gradient descent methods are not directly applicable. A method similar to gradient descent for minimizing convex non-differentiable functions is the *subgradient method*. Given an initial $\boldsymbol{\lambda}^0$ (e.g., $\boldsymbol{\lambda}^0 = \mathbf{0}$), it generates a sequence of Lagrange Multipliers $\{\boldsymbol{\lambda}^t\}$ by following the update rule

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t - \alpha_t \mathbf{g}^t$$

where $\alpha_t \in \mathbb{R}_{>0}$ is the step size at iteration t and $\mathbf{g}^t$ is a subgradient of $L(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^t$ that can be efficiently computed by solving the two subproblems in (6). The complete algorithm is given in Figure 2

The algorithm terminates either if in any iteration t the constraint $\mathbf{z}_h^t = \mathbf{z}_v^t$ is met or when the maximum number of iterations T has been reached. In the first case, an optimal joint gene structure $\mathbf{x}_{\text{exact}} = (\mathbf{y}_h^t, \mathbf{y}_v^t, \mathbf{z}_h^t)$ has been found. In the second case, a near optimal joint gene structure can be obtained as follows: In each iteration t, a potential joint gene structure $\mathbf{x}_p^t$ can be recovered from the dual solution. If $\mathbf{z}_h^t \neq \mathbf{z}_v^t$, i.e. when we have two inconsistent labelings of the exon candidates that are also leaf nodes in a tree, we chose to give precedence to the labeling $\mathbf{z}_h^t$, because it represents together with $\mathbf{y}_h^t$ biologically valid gene structures in each of the species. We therefore chose in line 10 the optimal ancestral labeling for the labeling $\mathbf{z}_h^t$ of the leaf nodes. The potential joint gene structure $\mathbf{x}_{\text{approx}} = \mathbf{x}_p^{t'}$, $t' = \text{argmax}_{t=0}^T S(\mathbf{x}_p^t)$ with highest score over all iterations is our best guess.

Choosing a good step size is crucial for convergence and speed of convergence. If the sequence of step sizes $\{\alpha_t\}$ is diminishing and non-summable, e.g.

$$\lim_{t\to\infty} \alpha_t = 0, \quad \sum_{t=0}^{\infty} \alpha_t = \infty$$

convergence of the dual problem is guaranteed (Nedic and Bertsekas, 2001). Thus

$$\lim_{t\to\infty} L(\boldsymbol{\lambda}^t) = \min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$$

The complexities of the longest path search and pruning algorithm are $\mathcal{O}(|E_D| + |V|)$[3] and $\mathcal{O}(|V|)$, respectively. Therefore, the runtime for determining an optimal or near optimal joint gene structure is proportional to $\mathcal{O}(T(|E_D| + |V|))$.

[3] In our implementation $|E_D| = O(|V|)$ although the number of intron candidates grows quadratically with the number of exon candidates. This is achieved by introducing at most two auxiliary nodes for each exon candidate.

```
$\boldsymbol{\lambda}^0 \leftarrow \mathbf{0}$ // initialization
$\mathbf{x}_{\text{approx}} \leftarrow \mathbf{0}$ // best approximative joint gene structure so far
for $t = 0, 1, \ldots, T$ do
    $(\mathbf{y}_h^t, \mathbf{z}_h^t) \leftarrow \text{argmax}_{\mathbf{y}_h, \mathbf{z}_h} h(\mathbf{y}_h, \mathbf{z}_h) + \boldsymbol{\lambda}^{t\top}\mathbf{z}_h$ // DAG-longest path
    $(\mathbf{y}_v^t, \mathbf{z}_v^t) \leftarrow \text{argmax}_{\mathbf{y}_v, \mathbf{z}_v} v(\mathbf{y}_v, \mathbf{z}_v) - \boldsymbol{\lambda}^{t\top}\mathbf{z}_v$ // pruning algorithm
    if $\mathbf{z}_h^t = \mathbf{z}_v^t$ then
        $\mathbf{x}_{\text{exact}} \leftarrow (\mathbf{y}_h^t, \mathbf{y}_v^t, \mathbf{z}_h^t)$
        return $\mathbf{x}_{\text{exact}}$
    else
        $\mathbf{x}_p^t \leftarrow (\mathbf{y}_h^t, \text{argmax}_{\mathbf{y}_v} v(\mathbf{y}_v, \mathbf{z}_h^t), \mathbf{z}_h^t)$ // potential joint gene structure
        if $S(\mathbf{x}_{\text{approx}}) < S(\mathbf{x}_p^t)$ then
            $\mathbf{x}_{\text{approx}} \leftarrow \mathbf{x}_p^t$
        end if
        $\boldsymbol{\lambda}^{t+1} \leftarrow \boldsymbol{\lambda}^t - \alpha_t(\mathbf{z}_h^t - \mathbf{z}_v^t)$ // subgradient update
    end if
end for
return $\mathbf{x}_{\text{approx}}$
```

Figure 2. The dual decomposition algorithm for finding an optimal or near optimal joint gene structure $\mathbf{x}$.

RESULTS

Datasets

We tested our method on a data set of 12 genomes of different *Drosophila* species (*D. melanogaster* (r6.04), *D. sechellia* (r3.03), *D. simulans* (r2.01), *D. yakuba* (r1.04), *D. erecta* (r1.04), *D. ananassae* (r1.04), *D. pseudoobscura* (r3.03), *D. persimilis* (r1.3), *D. willistoni* (r1.3), *D. mojavensis* (r1.3), *D. virilis* (r1.2), and *D. grimshawi* (r1.3)) available on FlyBase (http://flybase.org). All genomes were soft-masked with REPEATMASKER (Smit et al., 2015) using the standard Repbase *Drosophila* library and TRF (Benson, 1999). An alignment of the masked genomes was build with PROGRESSIVE CACTUS (Paten et al., 2011). For constructing a phylogenetic tree, orthologous Flybase genes were identified by reciprocal best BLAST hit comparison between *D. melanogaster* and each of the other *Drosophila* genomes. A random subset of 100 orthologs across all 12 genomes were selected and their nucleotide sequences including introns were aligned using CLUSTAL OMEGA (Sievers et al., 2011) followed by a manual trimming of alignment ends. Finally, a phylogenetic tree based on the concatenated alignments was obtained by inferring branch lengths on the known species tree topology (Stark et al., 2007) with FASTTREE (Price et al., 2010) using the GTR model with 20-gamma-distributed rate categories. Note, that the phylogenetic tree is both input of PROGRESSIVE CACTUS and AUGUSTUS$_{\text{CGP}}$.

We evaluated the accuracy of the predictions on *D. melanogaster* only, which has the most mature annotation. We thereby compared predictions on the *D. melanogaster* genome only with the FlyBase gene annotation. The filtering tool GENE-CHECK from the UCSC genome browser group was applied to remove questionable transcripts (e.g. with in-frame stop codon, splice site pairs other than GT-AG, GC-AG or AT-AC, missing start or stop codon or a CDS length not a multiple of 3) from the FlyBase gene set. The filtered FlyBase annotation contained 13 789 genes and 21 440 transcripts. The conventional accuracy measures sensitivity and specificity of the prediction on gene, exon and nucleotide level, were calculated using the EVAL package (Keibler and Brent, 2003). The evaluation was done on protein coding regions only (CDS), although AUGUSTUS predicted UTRs in the RNA-Seq-based experiments, as well. An exon is classified as correctly predicted if both its boundaries coincide with a FlyBase exon. A gene is counted as correct if it matches the coding region of one splice form of a FlyBase gene exactly.

De novo Performance

The comparison with the currently most accurate *de novo* gene predictors N-SCAN and CONTRAST has turned out to be very difficult, since both require a far from trivial parameter training specific to the set of input genomes. On these grounds, we restricted the comparison to N-SCAN that provides a parameter set that can be used for gene finding in *D. melanogaster* with *D. ananassae* as the only informant. According to previous results, *D. ananassae* was the best single informant with an optimal evolutionary distance from *D. melanogaster* (Brent, 2008) and N-SCAN actually never showed big improvements on fly when further informants were added (Michael Brent, personal communication, 2015). Note however that these findings are from 2008. Since then, the quality of the genome assemblies improved considerably and it is quite possible that N-SCAN performs better with multiple informants if retrained using more recent assemblies. Furthermore, we compare the standard version of AUGUSTUS that determines the most likely gene structure in a single input genome with the new comparative version AUGUSTUS$_{\text{CGP}}$. Although, AUGUSTUS$_{\text{CGP}}$ can predict exons that are not aligned, it is limited in so far as it cannot find genes in longer unaligned regions of a genome, e.g. a species-specific genomic region. To obtain a gene annotation of all whole genomes, the AUGUSTUS$_{\text{CGP}}$ gene set is merged with the AUGUSTUS gene set, giving a higher priority to the CGP version in the case of two conflicting versions of a gene. Both AUGUSTUS and AUGUSTUS$_{\text{CGP}}$[4] can report multiple transcripts per gene. However, for a more direct comparison with N-SCAN that determines only a single splice form per gene, alternative transcripts were discarded.

As shown in Table 1*a*, AUGUSTUS$_{\text{CGP}}$ is on all levels both more sensitive and specific compared to AUGUSTUS using a single genome and N-SCAN with a single informant. The accuracy values of N-SCAN are somewhat worse than the values reported in (Gross and Brent, 2006). It should be noted, however, that the FlyBase annotation had been revised in the mean time, containing several novel genes and a larger number of alternative splice forms - on average 1.55 per gene. This also explains some fraction of exons that are not predicted.

	gene Sn	gene Sp	exon Sn	exon Sp	nuc Sn	nuc Sp
(*a*) *de novo* methods						
AUGUSTUS	56.48	60.01	72.88	82.41	92.32	96.82
AUGUSTUS$_{\text{CGP}}$	63.10	64.86	76.37	85.22	95.37	97.67
N-SCAN	47.91	52.10	68.90	75.15	94.02	91.64
(*b*) AUGUSTUS with RNA-Seq						
	69.23	**74.93**	78.21	**89.95**	93.62	97.50
(*c*) AUGUSTUS$_{\text{CGP}}$ with RNA-Seq for						
D. mel	71.88	72.20	79.27	89.09	96.46	97.49
D. sim	67.10	67.43	77.64	86.67	95.96	97.71
D. mel + *D. sim*	74.33	73.13	80.22	89.74	96.77	97.57
all 4 *Drosophilas*	**74.46**	73.18	**80.31**	89.79	96.79	97.56
(*d*) cross-species annotation transfer from *D. sim* to *D. mel*						
AUGUSTUS$_{\text{CGP}}$	71.82	70.71	80.30	87.93	96.34	**97.77**
GENEWISE	34.76	20.01	67.28	65.05	**97.89**	96.76

Table 1. Sensitivity (Sn) and specificity (Sp) of the predictions in *D. melanogaster* (whole genome) at gene, exon and nucleotide (nuc) level (values are given in %). The evaluation was done on the coding regions (CDS) only. Whenever the subscript CGP is not used, AUGUSTUS refers to the standard version which here uses the *D. melanogaster* genome only. The genetic distances (expected number of mutations per genomic site) of *D. mel* to *D. sim*, *D. pse* and *D. vir* are 0.08, 0.71 and 1.03, respectively.

Performance with RNA-Seq Data

It is a good policy to combine information from many different sources of evidence. AUGUSTUS allows for integration of different types of extrinsic evidence including transcriptome data (RNA-Seq, cDNA, ESTs), protein sequences, and existing annotations. In AUGUSTUS$_{\text{CGP}}$ extrinsic evidence is species-specific and can be provided for each or a subset of the genomes. To see how well AUGUSTUS$_{\text{CGP}}$ performs with extrinsic evidence, we conducted several experiments incorporating RNA-Seq data for 1 to a maximum of 4 input genomes. Paired-end RNA-Seq reads were obtained from the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) and mapped to the corresponding (unmasked) genomes with STAR (Dobin et al., 2013). The resulting spliced alignments were filtered by coverage (minimum 80% of read length) and percentage identity (minimum 92%). If a read mapped to multiple locations, only the unique best alignment (in terms of coverage and percentage identity) for that read was kept, e.g. multiple almost equally best alignments were also discarded.

[4] in AUGUSTUS$_{\text{CGP}}$ alternative transcripts are rather a by-product of overlapping syntenic regions.

Table 2 shows how well the transcriptome of each of the 4 genomes is covered by its combined RNA-Seq libraries. Coverage was measured as the proportion of FlyBase transcripts that have at least 100 aligned reads per kilobase of mRNA.

	D. mel	*D. sim*	*D. pse*	*D. vir*	
Tx coverage	80.2	89.0	84.3	84.9	%

Table 2. Transcriptome coverage as measured by the proportion of FlyBase transcripts with at least 100 aligned reads per kilobase of mRNA.

With RNA-Seq evidence for *D. melanogaster*, AUGUSTUS$_{\text{CGP}}$ is slightly more sensitive compared to AUGUSTUS using the same evidence (compare Table 1c and Table 1b). When combining the *D. mel* evidence with the RNA-Seq evidence of the close *D. simulans*, we again observe a small boost in performance. However, beyond that, there are no significant improvements, when further adding RNA-Seq evidence of more distant flies (*D. pseudoobscura* and *D. virilis*). This could be explained by the fact that only a fraction of transcripts have RNA-Seq support, that were not already supported by RNA-Seq of *D. mel* or *D. sim*. Also, their distances to *D. mel* may be too large, to make any meaningful improvements upon an already good quality prediction. Note that such RNA-Seq data is still likely to improve the AUGUSTUS$_{\text{CGP}}$ accuracy on genomes close to *D. pseudoobscura* and *D. virilis*. We just did not evaluate the accuracy on these genomes. Using only RNA-Seq evidence of a non-target genome (*D. simulans*) still improves accuracy over *de novo* AUGUSTUS$_{\text{CGP}}$ predictions but produces poorer results that when using RNA-Seq from the target genome itself. This is generally to be expected, as the evidence can only be carried over to genes common to both the target and the informant, but not to genes that are exclusive to the target genome.

Liftover of Existing Annotations

An increasingly important strategy in genome annotation is the transfer of trusted annotations of previously existing genomes to newly sequenced genomes with a reasonable degree of sequence similarity, e.g. using protein spliced alignments Birney et al. (2004). AUGUSTUS$_{\text{CGP}}$ can be adapted for this purpose by compiling the existing annotations in a similar manner as other extrinsic data to intron and CDS 'hints'. We tested to what extent AUGUSTUS$_{\text{CGP}}$ succeeds in lifting over the FlyBase annotation of a non-*melanogaster* genome to *D. melanogaster*. Note that, in contrast to alignment-based strategies for annotation liftover, AUGUSTUS$_{\text{CGP}}$ can identify new genes and gene structures different from the source gene.

As shown in Table 1d, lifting over the filtered *D. simulans* FlyBase annotation to the target (*D. mel*) yields comparable levels of accuracy as using target-specific RNA-Seq evidence. Note, however, that the effectiveness of a liftover depends on the quality of the alignment and the annotation to be transferred. It is reasonable to assume that the *melanogaster* annotation is more comprehensive and accurate than the *simulans* annotation. Therefore in the typical use-case, where *D. mel* is the source and another genome is the target, the accuracy may be better than reported here for the reverse case.

An alternative approach to make use of an existing annotation is to use protein sequence spliced alignments rather than a genome alignment. For a comparison of the two approaches we ran GENEWISE Birney et al. (2004), a component of the ENSEMBL genebuild pipeline, in the following way. For each protein in the FlyBase reference gene set, that we use for evaluation, we ran BLASTP against the *D. sim* proteins from FlyBase and used the top 5 hits as well as the genomic region of the target *D. mel* transcript with 10 000bp padding on both sides as input to GENEWISE (options `-alg 623S` and `-splice_gtag`). Note that in a typical annotation setting for new genomes, the 'correct' target regions on which to run GENEWISE would have to be found using some fast alignment approach. We chose to give GENEWISE this conceptual advantage in order to save run-time and to mask out errors from target region detection.

Effectiveness of Dual Decomposition

Following step size used in previous applications (Koo et al., 2010), has shown to be most efficient

$$\alpha_t = \frac{c}{\sqrt{d+1}} \tag{7}$$

where $c \in \mathbb{R}_{>0}$ is a parameter that can be trained and $d < t$ is the number of iterations prior to the current, in which the value of the dual problem increases, e.g. $L(\boldsymbol{\lambda}^d) > L(\boldsymbol{\lambda}^{d-1})$. As a result, the step size only decreases when the dual problem moves in the wrong direction. In total, dual decomposition was applied to 4443 syntenic regions with an average size of 84 kbp in *D. melanogaster*. In 94% of the cases an exact solution was found (on average after 200 iterations). In all other cases the approximation error $\varepsilon := \min_{t=0}^{T} L(\boldsymbol{\lambda}^t) - S(\mathbf{x}_{\text{approx}})$ was less than 0.05% of the initial error $\varepsilon^0 := L(\boldsymbol{\lambda}^0) - S(\mathbf{x}_p^0)$ when stopping after 2500 iterations.

The running time (sum of CPU times over all threads) for *de novo* gene finding and cross-species annotation transfer was about 20 CPU days on a compute cluster with Intel Xeon E5440 processors (2.83 GHz). Gene finding with RNA-Seq evidence was generally more expensive (around 28 CPU days) as we used a model that includes untranslated regions (UTRs) in these cases. For comparison, annotating a single genome (*D. mel*) with N-Scan and GeneWise required 18 CPU hours and 10 CPU days, respectively. When splitting the genome alignment into smaller chunks of at most 100 MB, memory usage of Augustus$_{\text{CGP}}$ was below 4 GB for each alignment chunk.

DISCUSSION AND OUTLOOK

In this paper we have presented a novel approach to comparative gene finding that is suitable for the gene structure annotation of entire clades. Its novelty is that it simultaneously identifies genes in multiple genomes. Previous gene finding systems were either limited to exactly two genomes or restricted the prediction and gene structure model to a single target genome. Unlike the target-informant approach that requires a repetitive training of parameters for each and every genome to be annotated, parameters only have to be trained for a single representative in the clade. Beside the coding region (CDS) of a gene, our approach can also predict the UTR. This is particularly useful when incorporating RNA-Seq evidence that gives unspecific hints about both coding and non-coding parts of genes.

As a tendency, our approach favors gene structures that are in agreement across the genomes. Thus, it is likely to produce more consistent gene sets than the ones obtained from the individual annotation of each genome. This is particularly important when the objective of study is to investigate the genomic differences of several species within a clade.

The results show that the new multi-species version of Augustus is more accurate than the standard single-species version. In the *de novo* category where only genome evidence is used it compares favorably with N-Scan. In evidence-based gene finding our findings are, that when having RNA-Seq evidence for the target genome itself, there is very little additional benefit from RNA-Seq evidence from other species in the fly clade. This may, however, be different for other clades and libraries.

Annotation can be transfered using Augustus$_{\text{CGP}}$ from one genome to another via the multiple genome alignment. A previously existing option to do this is the alignment of the source proteins (or transcripts) to the target genome. Genome alignments however have the advantage that the context around the exons and genes is also used to identify what is homolog. For example, initial coding exons can be very short and therefore very difficult to align correctly in a protein alignment, even if the genomes are similar, whereas a genome alignment may have no difficulty when the neighboring UTR or intron is also alignable. At very large distances, however, where genome alignments are hardly or not at all possible, protein (family) homology searches are still useful, at least to identify conserved domains of the gene.

The dual decomposition approach has proven to be a well-suited framework to efficiently obtain good approximate and even mostly exact solutions to the formal optimization problem of comparative gene finding.

A common weakness of gene predictors is to distinguish between correct candidate exons and partly correct candidate exons that only differ from each other by a few base pairs. When, for example, classifying all exons as correctly predicted that overlap a true exon by at least 80% of the length of the longer one, Augustus$_{\text{CGP}}$ achieves *de novo* an exon sensitivity of 89.94%. In other words, around $(89.94\% - 76.37\%)/(100\% - 76.37\%) \approx 57\%$ of the false negative FlyBase exons are close to correctly predicted. We continue to work on the vertical scoring function in order to improve the precision of exon boundary prediction using the multiple genome alignment.

REFERENCES

Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM: Cross-species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model. *Genome Res.*, 13(3):496–502.

Althaus, E. and Canzar, S. (2008). A lagrangian relaxation approach for the multiple sequence alignment problem. *J. Comb. Optim.*, 16(2):127–154.

Andreotti, S., Klau, G. W., and Reinert, K. (2012). Antilope - a lagrangian relaxation approach to the de novo peptide sequencing problem. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 9(2):385–394.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.*, 14:988–995.

Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet.*, 9(1):62–73.

Csűrös, M. (2006). On the estimation of intron evolution. *PLoS Comput. Biol.*, 2(7):e84.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, (1):15–21.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates.

Genome 10K Community of Scientists (2009). Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, 100(6):659–74.

Gross, S. S. and Brent, M. R. (2006). Using Multiple Alignments to Improve Gene Prediction. *J. Comp. Biol.*, 13(2):379–393.

Gross, S. S., Do, C. B., Sirota, M., and Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.*, 8(12):R269+.

Hoff, K. and Stanke, M. (2015). Current methods for automated annotation of protein-coding genes. *Current Opinion in Insect Science*.

Keibler, E. and Brent, M. R. (2003). Eval: A software package for analysis of genome annotations. *BMC Bioinform.*, 4:50.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, 27(6):757–763.

Komodakis, N., Paragios, N., and Tziritas, G. (2011). MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552.

Koo, T., Rush, A. M., Collins, M., Jaakkola, T., and Sontag, D. (2010). Dual decomposition for parsing with non-projective head automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1288–1298, Stroudsburg, PA, USA. Association for Computational Linguistics.

Meyer, I. M. and Durbin, R. (2002). Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, 18(10):1309–1318.

Nedic, A. and Bertsekas, D. P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM J. Optim.*, 12(1):109–138.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.*, 21(9):1512–1528.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J., Robertson, H. M., and Schneider, D. J. (2011). Creating a buzz about insect genomes. *Science*, 331(6023).

Rush, A. M., Sontag, D., Collins, M., and Jaakkola, T. (2010). On dual decomposition and linear programming relaxations for natural language processing. In *In Proc. EMNLP*.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 7(1).

Smit, A. F. A., Hubley, R., and Green, P. (2013-2015). RepeatMasker Open-4.0.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cdna alignments to improve *de novo* gene finding. *Bioinformatics*, 24(5):637–644.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). Augustus: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34:435–439.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., et al. (2007). Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232.

Steijger, T., Abril, J., Engstrom, P., Kokocinski, F., Consortium, T. R., Hubbard, T., Guigo, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, 10(12):1177–1184.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1296v1 | CC-BY 4.0 Open Access | rec: 11 Aug 2015, publ: 11 Aug 2015