BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 5 2008, pages 637–644
doi:10.1093/bioinformatics/btn013

Sequence analysis

Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding

Mario Stanke*, Mark Diekhans, Robert Baertsch and David Haussler

Center for Biomolecular Science and Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

Received on October 9, 2007; revised on December 12, 2007; accepted on January 7, 2008

Advance Access publication January 24, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Computational annotation of protein coding genes in genomic DNA is a widely used and essential tool for analyzing newly sequenced genomes. However, current methods suffer from inaccuracy and do poorly with certain types of genes. Including additional sources of evidence of the existence and structure of genes can improve the quality of gene predictions. For many eukaryotic genomes, expressed sequence tags (ESTs) are available as evidence for genes. Related genomes that have been sequenced, annotated, and aligned to the target genome provide evidence of existence and structure of genes.

Results: We incorporate several different evidence sources into the gene finder AUGUSTUS. The sources of evidence are gene and transcript annotations from related species syntenically mapped to the target genome using TransMap, evolutionary conservation of DNA, mRNA and ESTs of the target species, and retroposed genes. The predictions include alternative splice variants where evidence supports it. Using only ESTs we were able to correctly predict at least one splice form exactly correct in 57% of human genes. Also using evidence from other species and human mRNAs, this number rises to 77%. Syntenic mapping is well-suited to annotate genomes closely related to genomes that are already annotated or for which extensive transcript evidence is available. Native cDNA evidence is most helpful when the alignments are used as compound information rather than independent positionwise information.

Availability: AUGUSTUS is open source and available at http://augustus.gobics.de. The gene predictions for human can be browsed and downloaded at the UCSC Genome Browser (http://genome.ucsc.edu)

Contact: mstanke@gwdg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Eukaryotic genome projects are dependent on automated methods for identifying the genes in the genome assembly. However, the accuracy of current methods is not sufficient for many purposes. For example, taking alternative splicing into account, the best methods predict only 40–50% of human transcripts correctly (Guigó *et al.*, 2006). Increasing the prediction accuracy is crucial to ensuring a maximal scientific benefit of the genome projects. So-called *ab initio* gene finders just require the target genome as input and some are readily trained for a new genome (Korf, 2004; Lomsadze *et al.*, 2005; Stanke and Waack, 2003) However, the progress of the last years in improving *ab initio* methods is not far reaching enough (Guigó *et al.*, 2006) and the methods of choice in most genome projects are methods based on extrinsic data such as expressed sequence tags (ESTs), other cDNA evidence, and the information that is provided by related sequenced genomes and their annotation. However, *ab initio* methods are still required to find genes without sufficient support by extrinsic data.

Many genome sequencing projects are also accompanied by EST sequencing efforts. The new pyrosequencing (454) method (Margulies *et al.*, 2005) is also resulting in an increase in the availability of cDNA reads. Several methods have been developed that predict genes based on full-length coverage with transcript alignments and/or protein alignments (e.g. Djebali *et al.*, 2006; Thierry-Mieg and Thierry-Mieg, 2006). However, even with exceptionally extensive cDNA sequencing, as in the case of the human genome, not all genes are covered with transcript alignments. Further, many genes are only partially covered with ESTs and the remaining part needs to be predicted without the help of any cDNA evidence.

Gene finders which incorporate ESTs in their predictions can improve their accuracy over pure *ab initio* predictions (Brejová *et al.*, 2005; Stanke *et al.*, 2006b). In a recent study, the program N-SCAN/EST, which is based on an *ab initio* model, has been shown to predict approximately 40% of human genes correctly using only ESTs and conservation to three other vertebrate genomes (Wei and Brent, 2006). However, N-SCAN/EST does not incorporate other extrinsic information, it does not take alternative splicing into account, and it does not predict the 3′UTR of genes, which is often covered best by ESTs. Another gene finder that can incorporate ESTs and that does not completely rely on cDNA data for the prediction is EuGÈNE-M (Foissac and Schiex, 2005). It constrains the gene structure in areas covered by transcripts and uses an *ab initio* model for the prediction in the other areas and for the identification of the coding start and stop. However, the extrinsic

*To whom correspondence should be addressed.

information that is incorporated is limited to transcript alignments. Further, it is trained only for a few plant species and therefore the current version of the program cannot be properly applied to vertebrate genomes.

There are a number of dual or multiple genome-based gene finders which use unannotated related genome sequences to predict the genes in one or two of these sequences by exploiting that functional regions are evolutionary better conserved (Gross and Brent, 2005; Meyer and Durbin, 2002). However, when a new genome is being sequenced, increasingly, related sequenced genomes that previously exist already have an annotation or cDNA data. This is particularly the case for the new mammalian genome projects which can make use of conservation with other mammals and the extensive annotations and mRNA data of human, mouse, rat and other mammals. For this reason, methods have been developed which explicitly exploit the information that the annotation from a syntenic genome sequence provides for annotating the genes in a target genome. Projector (Meyer and Durbin, 2004) uses a pair of homologous sequences from two species to project known genes from one species to the other using a pairHMM. However, this cannot easily be generalized to using multiple informant species and Meyer and Durbin did not propose an annotation pipeline for entire genomes. Another method is used in the AIR pipeline (Florea *et al*., 2005) which uses a splice graph to find and weight transcripts based on evidence from cDNA and protein sequences. In particular, AIR incorporates annotated features mapped from a closely related species using syntenic mapping information.

In many species, alternative splicing is too frequent to be neglected in the annotation process. There exist a number of non-expression-based methods that allow predicting more than one splice form per gene. The approaches are to predict suboptimal gene structures (HMMGene, Krogh, 1997) or sampling-based (Cawley and Pachter, 2003; Stanke *et al*., 2006a). The program ExAlt (Allen and Salzberg, 2006) takes a given transcript as input and tries to find alternative splicing, explicitly modeling certain alternative splicing events. Currently, the non-expression-based methods are not accurate enough for a reliable large-scale annotation of alternative splice forms and the most reliable way to predict alternative splicing requires actual transcript evidence for the splice form. Expression-based annotation pipelines that report alternative splicing are ENSEMBL (Curwen *et al*., 2004), Aceview (Thierry-Mieg and Thierry-Mieg, 2006) and EuGÈNE-M.

The *ab initio* version of the General-Hidden-Markov-Model-based (GHMM) gene finder AUGUSTUS has been shown to be among the most accurate for several species (Brejová, 2005; Guigó *et al*., 2006; Korf, 2004). In this article, we present an extension of AUGUSTUS, which can incorporate evidence from a variety of sources, predict alternative splicing and allows predicting complete genes including both UTRs and introns therein. This gene finder has a significantly higher accuracy than existing systems when only ESTs are available or only ESTs and related genome sequences are available. Further, we exploit mRNA alignments of other genomes through using the syntenic alignment to the target genome. Our system is capable of incorporating extrinsic information from conservation, native EST and mRNA alignments as well as alien transcript alignments. This enables us to predict up to 77% of genes correctly on the human genome when using all information at the same time. The presented method is very general and allows users to provide evidence for a gene structure from other sources of extrinsic evidence as input.

2 METHODS

2.1 Incorporating hints from extrinsic evidence

Extrinisic evidence is computed or collected beforehand and given as input to AUGUSTUS in the form of 'hints' in a file in GFF format. With the word *hint* we are referring to an uncertain local piece of information about the gene structure of the input sequence such as a likely position of a signal or a likely stretch of coding sequence. We distinguish 16 types of hints shown in Table 1. Each type is associated with a biological label of the gene structure as indicated by the name of the type. Each hint specifies an interval or a position of the target DNA sequence. The hints may also contain strand information

Table 1. Types of hints

Type *t* of hint	Description
start	Translation start
stop	Translation stop
tss	Transcription start site
tts	Transcription termination site
ass	Acceptor (3′) splice site
dss	Donor (5′) splice site
exon	Exact exon
exonpart	Part of an exon
intron	Exact intron (in CDS or UTR)
intronpart	Part of an intron
CDS	Coding part of an exon with exact boundaries
CDSpart	Part of the coding part of an exon
UTR	Exact boundaries of a UTR exon or the untranslated part of a partially coding exon
UTRpart	Part of a UTR
irpart	Part of the intergenic region
nonexonpart	Part of intergenic region or intron

'Exon' refers here to exons in the biological sense, i.e. an exon can be completely coding (CDS), completely untranslated (UTR) or it can be partially coding and partially untranslated.

Table 2. Types of hints we used depending on source of evidence

Source of extrinsic information	Types of hints
ExoniPhy	CDS, CDSpart
PhastCons	CDSpart
native ESTs, mRNAs	exon, intron, exonpart
TransMap RefSeqs	CDSpart, intronpart, intron, UTRpart, start, stop, tss, tts
TransMap mRNAs without ORF information	exonpart, intronpart, intron, tss, tts
retroposed genes	nonexonpart

and a reading frame, if appropriate. Table 2 shows which types of hints we generate from each source. In a given application setting with various heterogenous sources of available extrinsic evidence, all hints of all available sources are used simultaneously. The files containing the hints for each source are simply concatenated.

We will use the following vocabulary. By a *single-transcript gene structure* we refer to a gene structure in which every gene has only one transcript and no genes are overlapping. A single-transcript gene structure *obeys* a hint if the gene structure has the biological feature as specified in the hint. In that case, we also say that the hint *supports* the gene structure. Two hints are said to be *compatible* if there exists a single-transcript gene structure which obeys both hints. A hint g is said to *support* a hint h if every single-transcript gene structure which obeys g obeys also h. For example, if g and h are CDSpart hints in the same frame and on the same strand, then g supports h if and only if the interval of g contains the interval of h.

Consider a fixed set of sources of evidence, a single type t and the biological feature (label) f that type of hints refers to. For each type, we define a penalty factor $0 \leq m(t) \leq 1$, that we call *malus*. When a candidate single-transcript gene structure is evaluated by the GHMM, each feature f of the gene structure that is not supported by any hint is penalized by multiplying the factor $m(t)$ to its probability. For the types ending in 'part', the malus is applied for each base position of the exon, intron, or intergenic region that is not covered by the hint. For all other types it is applied only once for each signal or interval. Further, with each hint h a bonus factor $b(h) > 1$ is associated. A candiate gene structure is rewarded for each feature f by multiplying its probability with

$$\prod_h b(h),$$

where h ranges over all hints that support feature f. For example, consider *intron* hints. Then, for each hint h indicating an intron, that intron is rewarded by multiplying its probability with $b(h)$. An intron candidate for which no supporting intron hints exists is penalized by multiplying its probability with $m(t)$.

The bonus factor $b(h)$ depends mainly on the type of hint and the source of evidence, but can also depend on a score associated with the hint and the degree of compatibility of the hint with all other hints. Let H be the collection of all hints. Then

$$b(h) := [b(t, s) \cdot b(\text{score(h)})]^{c(h, H)}.$$

$b(t, s)$ is a bonus factor associated with the type t and the source s, taking into consideration the fact that some sources are more reliable than others. $b(\text{score}(h))$ is a function of the score field in the GFF file. For almost all types and sources in our applications presented in this paper this is not used (i.e. $b(\text{score}(h)) = 1$). For the other cases, we distinguish only 'low'- from 'high'-scoring hints. $0 < c(h, H) < 2$ is a modifier depending on how well h is compatible with other hints in H: $c(h, H) = 2(S + 5)/(S + I + 10)$, where S is the number of hints in H that support h and I is the number of hints in H that are incompatible with h. So, when there are few other hints, supportive or incompatible, $c(h, H)$ is close to 1 and does not affect the bonus much. $c(h, H) < 1$ if and only if there are more incompatible hints than supportive. We introduced this modification to downweight the impact of spurious EST alignments that often disagree with the majority of other EST alignments. In our experiments, ESTs were by far the source with the highest number of hints.

The number of parameters for the integration of hints is relatively small, often small enough to be adjusted by hand. For example, in a *de novo* setting, where only hints from ExoniPhy (Siepel and Haussler, 2004) are used, there are only four parameters. We use malus(CDS) = 0.25, b(CDS, ExoniPhy) = 25, malus(CDSpart) = 0.992, b(CDSpart, ExoniPhy) = 400. For comparison, N-SCANs fifth-order Markov chain for the conservation sequence requires the training of several thousands of parameters. In the most inclusive application setting, when hints from ExoniPhy, TransMap Refseqs, ESTs, mRNAs and retroposed genes are used at the same time, the total number of parameters for the hints is 29. These parameters for the hints are listed in the configuration files extrinsic.*.cfg in the supplementary data.

2.2 Alternative splicing

Each individual hint contains only local information. However, in the presence of alternative splicing or, in general, when the transcript alignments 'contradict' each other, an alignment usually contains more information than the set of independent hints derived from it. It also indicates that these hints belong to the same transcript. Therefore, we allow that the hints are *grouped* such that all hints from one group are thought to belong to the same transcript. All hints that are derived from a single (native or TransMap) alignment are grouped together. Hints from different alignments belong to different groups. Similarly, the CDSpart and CDS hint derived from a single predicted ExoniPhy exon belong to one group.

In an initial step of the AUGUSTUS algorithm the compatibility of all hint groups with each other is determined. Two hint groups are considered compatible if and only if there is a possible single-transcript gene structure that is compatible with both hint groups. In particular, alignments suggesting alternative splice variants of a gene yield incompatible hint groups. Also, alignments suggesting a gene contained in an intron of another gene or overlapping genes on opposite strands yield incompatible hint groups. The input sequence is then internally split into segments such that no hint groups are separated and no genes are separated, if possible. Here, we use predicted genes from a first run of the Viterbi algorithm using all hints simultaneously.

In a subsequent step we make several prediction runs for each segment, each using the Viterbi algorithm. In each prediction run, a different subset of the hint groups is deactivated, meaning that these hints are not used at all in the prediction run. The set of prediction runs is the smallest set of prediction runs such that for each hint group g there is at least one prediction run in which all groups that are incompatible with g are deactivated. This gives AUGUSTUS the chance to predict splice forms that are compatible with *all* the hint groups. However, if the bonuses and maluses are moderate, AUGUSTUS may not obey all hint groups, since it is allowed to ignore individual hints. This will happen if all gene structures that obey a certain hint have, even with the bonus, a lower likelihood than some gene structure that disobeys the hint. Figure 1 shows an example of a prediction obeying some hints for alternative splicing, but not all.

In a post-processing step, transcripts from the previous step are grouped to genes. For each transcript, the fraction of coding exons, non-coding exons, and introns is computed that is supported by hints (see Fig. S1). Transcripts whose supporting fraction is smaller than 86% of the supporting fraction of another transcript of the same gene are discarded.

2.3 Pre-processing of hints

Hint groups from mRNA or EST alignments are deleted by AUGUSTUS if any of the hints in the group is unsatisfiable. Introns are required to have the GT/GC-AG consensus. (AUGUSTUS predicts the much less frequent GC-donor splice site only when it is supported by a hint.) Some loci of the human genome contain thousands of EST alignments and therefore ususally also a very large number of incompatible hint groups solely due to errors. To reduce the number of prediction runs triggered by spurious EST alignments, AUGUSTUS further discards hint groups where a hint has a suspiciously high number I of incompatible other hints compared to the number S of other hints supporting it. We discard the group if $I \geq 9(S + 1)$.

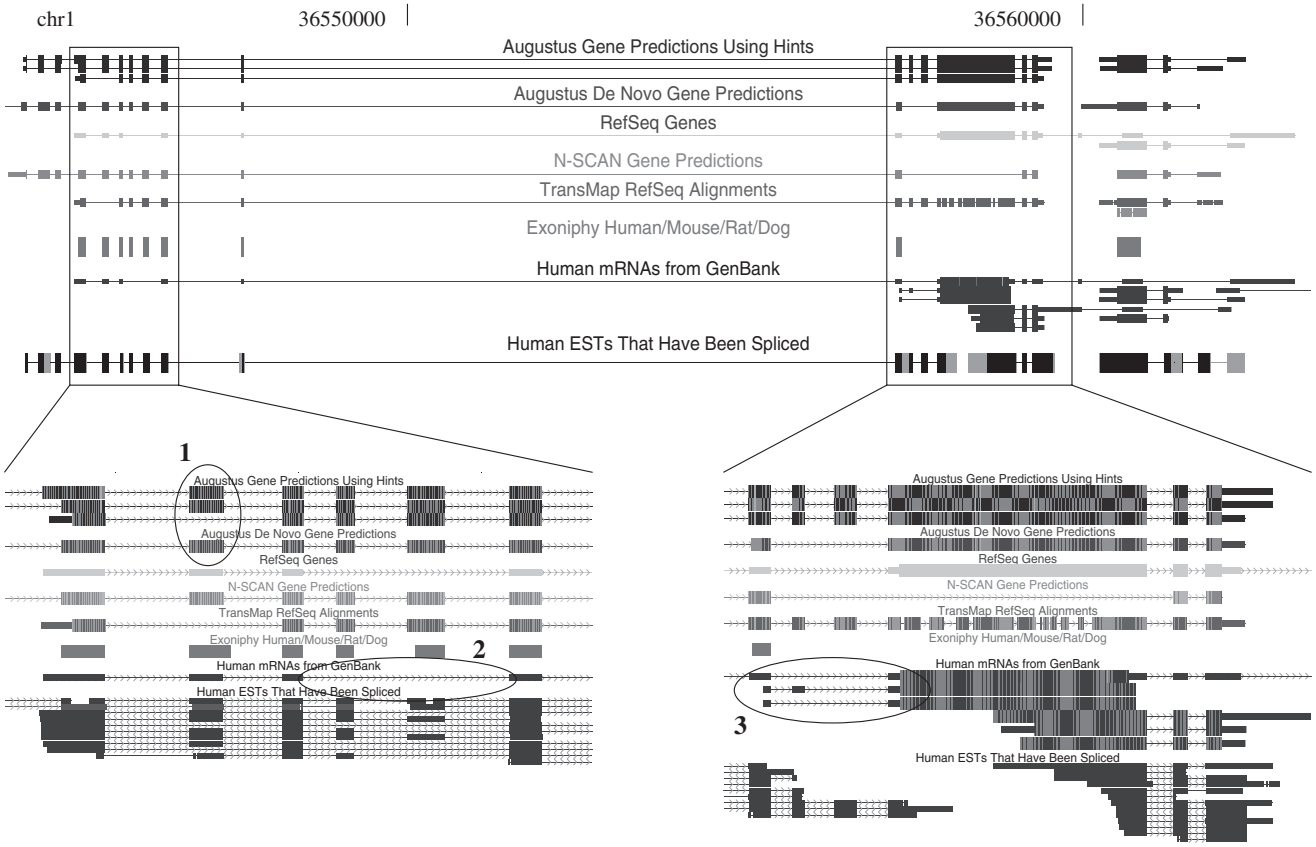

Fig. 1. Example annotation of a human region (assembly hg18) taken from the UCSC Genome Browser (Kuhn *et al*., 2006). The top AUGUSTUS track (= X,R,T,E,m) uses, in particular, the information from the other shown tracks TransMap, ExoniPhy, mRNAs and ESTs. In the circled region labeled 1 AUGUSTUS predicts alternative splicing supported by the EST alignments. In the circled regions 2 and 3 the mRNA track is missing exons which are supported by ESTs, a TransMap mouse RefSeq and cross-species conservation (circle 2 only) and predicted by AUGUSTUS. Here, a purely mRNA-based gene prediction contradicts most other evidence because of ORF constraints it is forced to predict most exons as untranslated (see RefSeq track).

The drawback is that splice variants supported by less than 10% of the EST evidence might not be found.

Hint groups can be given a priority number. In this case, any hint group which is incompatible with another hint group of higher-priority is discarded. The purpose of this procedure is to use not-so-reliable hints only when no better information source is available for the gene. For example, native transcripts can be set to override alien transcripts to account for divergence in the gene structures. However, parts of the gene structure suggested by higher priority hint groups containing incomplete information can be extended using lower-priority hint groups. In our experiments, we used the following priorities depending on the source of the hints: retrogenes 5, human RNAs 4, TransMap alignments 4, ExoniPhy 4, human spliced ESTs 2.

2.4 Generation of hints

TransMap *Orthologous Genes* Orthologous gene hints were produced by TransMap (Siepel *et al*., 2007; Zhu *et al*., 2007), a methodology for generating cross-species genomic alignments of cDNAs by combining the results of two alignment methods that are optimized for different tasks. Alignments of cDNA sequences to their cognate genome are done using BLAT (Kent, 2002). BLAT is designed to align transcripts of at least 95% identity to DNA sequences, producing intron-spanning alignments of the full cDNA. TransMap projects the cDNA to genome alignments through BLASTZ (Schwartz *et al*., 2003) cross-species genomic alignments to a target species genome. BLASTZ is a highly-sensitive aligner, optimized for aligning diverged, orthologous genomic sequences. BLASTZ alignment chains identified as syntenic (Kent *et al*., 2003) are used in the mappings. The use of alignment chains allows for the mapping of mRNAs as a whole, rather than as independent exons. The syntenic filtering removes paralogs, with the exception of those caused by tandem gene duplication, and alignments to processed pseudogenes.

TransMap was chosen for generating orthologous gene hints due to its relative immunity to pseudogenes and the ability to map a larger fraction of the gene structure than protein translation alignment methods. For a detail description of the TransMap methodology, see Zhu *et al*. (2007). The cDNA alignments produced by TransMap are used to generate hints, without the heuristic correction of gene structure used in this work. The evolutionary changes in gene structure were handled by generating hints that are *part* hints, which do not delineate the exact beginning and end of features. Alignment gaps that correspond to the location of introns in the mRNA in the source organism are used to generate intron hints.

The TransMap hints were generated by mapping the BLAT alignments of RefSeq mRNAs (Pruitt *et al*., 2007) for five organisms obtained from the UCSC genome browser database (Kuhn *et al*., 2006). The number of transcripts that were mapped from mouse, rat, cow,

Table 3. Accuracy on the ENCODE test regions against the reference annotation from EGASP

Program	Gene		Transcript		Exon		Base		tr./ gene
	sn[%]	sp[%]	sn[%]	sp[%]	sn[%]	sp[%]	sn[%]	sp[%]	
De novo methods									
AUGUSTUS +X,p	33.78	37.04	16.02	37.04	66.39	82.99	84.06	88.02	1
AUGUSTUS +X,p,RA	32.09	39.26	15.25	39.26	65.45	84.41	80.47	90.22	1
N-SCAN*	35.47	36.71	16.95	36.71	67.66	82.05	85.38	89.02	1
Methods using only genome sequences and ESTs									
AUGUSTUS +E	56.76	53.16	30.82	39.36	77.35	80.98	86.74	87.86	1.6
AUGUSTUS +X,p,E	53.38	59.40	26.50	49.85	75.80	86.57	84.27	92.75	1.27
AUGUSTUS +X,p,E 1tr	51.01	58.08	24.35	58.08	74.24	88.22	82.10	93.51	1
N-SCAN/EST	36.15	35.55	17.41	35.55	72.00	84.14	84.46	91.18	1
Methods using any type of information									
AUGUSTUS +X,R,T,E,m	77.36	72.96	46.84	48.70	84.99	82.36	94.89	90.88	2.0
AUGUSTUS +X,R,T,E,m 1tr	67.57	62.89	32.05	62.89	80.39	87.34	92.94	91.91	1
JIGSAW*	72.64	65.95	34.05	65.95	80.61	89.33	94.56	92.19	1
ENSEMBL*	71.62	67.32	39.75	54.64	77.53	82.65	90.18	92.02	1.47
N-SCAN PASA-EST	58.89	59.65	34.96	39.62	75.55	82.43	87.28	88.22	2.01
Methods using no transcribed data of target species									
AUGUSTUS +T	62.16	65.70	30.82	57.02	74.93	87.49	89.64	90.87	1.23

Sources of information: X: ExoniPhy, p: PhastCons, RA: retroposed genes based on AUGUSTUS prediction, R: retroposed genes based on mRNA, T: TransMap RefSeqs from other species, E: human ESTs, m: human mRNAs. tr./gene: average number of distinct transcripts (CDS only) per gene. *These accuracy values were taken from Guigó *et al.* (2006).

chicken and dog were 18887, 8976, 7323, 3095 and 848, respectively (see also Table 1). See Figure 2 for an example. The TransMap alignments were converted to hints of types *intron*, *CDSpart*, *UTRpart*, *start*, *stop*, *tss*, *tts* and *intronpart* with the script `transmap2hints.pl` in the AUGUSTUS package.

2.4.1 Processed pseudogenes To identify and analyze the functionality of relatively recently evolved retrogenes, we carried out BLASTZ alignments of a set of mRNAs against the human genome and then scored a set of features indicative of such retroposition. These features include the number of processed introns; the absence of conserved splice sites; breaks in orthology with mouse, dog, and rhesus monkey; the presence, position, and length of the poly (A) tail; and sequence similarity and fraction of the parent mRNA that is represented in the retrogene, indicating evidence of the likelihood of retroposition. As mRNA set we either used the actual human mRNA sequences (in the setting AUGUSTUS +R) or the set of predicted mRNAs (in the setting AUGUSTUS +RA), which have previously been predicted with *de novo* AUGUSTUS. The latter we did to simulate the absence of mRNA sequence data for the target genome.

2.4.2 cDNA alignments The mRNA and spliced EST alignments were taken from the UCSC Genome Browser tracks. They had been constructed using BLAT (Kent, 2002). Hints of the types *intron*, *exon* and *exonpart* were generated from the BLAT output with the script `blat2hints.pl` that comes with the AUGUSTUS distribution.

2.4.3 Conservation Each conserved coding exon predicted by ExoniPhy (Siepel and Haussler, 2004) gives rise to one hint of type CDS and one hint of type CDSpart. The CDSpart intervals are cut off by 9bp on both sides with respect to the ExoniPhy intervals. The reason for using both a CDS hint with exact boundaries and a CDSpart hint is that ExoniPhy exons are often correct but in the cases when they are not correct, they are often still approximate to a real exon. Further, in the *de novo* category, we used hints of type CDSpart created from the PhastCons conserved elements predictions (Siepel *et al.*, 2005) using the script `phastconsDB2hints.pl` in the AUGUSTUS distribution.

2.5 Training and testing

AUGUSTUS has previously been trained on 1284 human genes retreived from Genbank in 2002. The parameters for the hints were trained to optimize prediction accuracy against RefSeq genes on a part of human chromosome 17, which has no overlap with the ENCODE regions. This was done using a simple semi-automatic optimization procedure, iterating changes to the bonuses and maluses and measuring accuracy against the RefSeq annotation. However, a fully manual adjustment of the hint parameters to new sources or species is similarly effective.

As a reference set for evaluation of gene predictions we used the ENCODE reference annotation on human genome version hg17 from the GENCODE consortium (Harrow *et al.*, 2006). Many gene finders have been evaluated on this set previously (Guigó *et al.*, 2006). The N-SCAN/EST predictions were taken from the UCSC Genome Browser. The N-SCAN/EST+PASA predictions were only available for the assembly hg18. We compared them against the ENCODE reference annotation mapped to assembly hg18.

3 RESULTS

We evaluated the gene predictions of AUGUSTUS using hints from different combinations of sources of evidence on the human ENCODE regions. In Table 3, we compare the accuracy to that of several other programs, in particular to some which had performed among the best in the ENCODE Genome Annotation Assessment Project (EGASP) (Guigó *et al.*, 2006).

It should be noted that it is hard to compute very precise absolute accuracy numbers as the reference annotation is likely to have errors as well. Nevertheless, we think it is appropriate for a relative comparison of different programs, especially when they only use a subset of the complete information.

Among the *de novo* programs, which use only the target genome and other genome sequences as input, N-SCAN previously was by far the best-performing program for human (Guigó *et al.*, 2006). Here we report the original EGASP results as the accuracy of the whole-genome predictions of N-SCAN on the UCSC Genome Browser are somewhat worse. In this category of programs, we gave AUGUSTUS hints from genomic conservation through ExoniPhy and PhastCons (AUGUSTUS +X,p). In another run (AUGUSTUS +X,p,RA), we also added hints from *de novo* predicted retroposed genes. Incorporating the hints from retroposed genes, removed a small percentage of predicted genes and increased the specificity of AUGUSTUS at the cost of a somewhat lower sensitivity. The impact of the retroposed gene predictions on the average accuracy is relatively small. These results are in agreement with van Baren and Brent (2006). However, these hints are helpful when trying to avoid false positive 'new' genes.

Next, we tried variants that use no other evidence than genome sequences and ESTs (no full length mRNAs). We tried a variant of AUGUSTUS that uses human ESTs only (AUGUSTUS +E), and two variants that use human ESTs in addition to genomic conservation (AUGUSTUS +X,p,E and AUGUSTUS +X,p,E 1tr). We compared to N-SCAN/EST, which has previously been reported to perform best on human whithin this category of programs (Wei and Brent, 2006). Using ESTs only, AUGUSTUS predicts 77% more transcripts correctly than N-SCAN/EST, while at the same time having a higher transcript specificity. Incorporating information from genomic conservation substantially improves the specificity over pure EST-based predictions. We also tried reporting only one transcript per gene (AUGUSTUS +X,p,E 1tr) for a more direct comparison to N-SCAN/EST. As can be expected, this generally further increases specificity so that more than 58% of the transcripts predicted by AUGUSTUS match perfectly a reference transcript (N-SCAN/EST: 36%).

We also compared programs that use human mRNAs as well as other evidence. In this category we gave AUGUSTUS evidence from human mRNAs; mouse, rat, cow, chicken and dog RefSeqs mapped to the human genome using TransMap; and genomic conservation and predicted retroposed genes using human mRNA. When AUGUSTUS is configured to predict alterntive splice variants (AUGUSTUS +X,R,T,E,m) it predicts at least one splice variant correctly for more than 77% of the reference genes. In the EGASP experiment, JIGSAW had, with 72.64%, the highest gene-level sensitivity. However, JIGSAW has in most measures a higher specificity, which, in part, can be explained by the fact that it only predicts one splice form per gene. N-SCAN PASA-EST is a recently developed combination of N-SCAN/EST with PASA (Haas *et al.*, 2003), in which PASA is used to create the input *ESTseq* for N-SCAN/EST from alignments of mRNAs and ESTs and also to find alternative splice variants to the single-transcript genes predicted by N-SCAN/EST.

To simulate the performance on a genome for which no transcribed data at all exists, but where the gene annotation of related genomes is available, we ran AUGUSTUS using only hints from running TransMap on RefSeqs from non-human species (mouse, rat, cow, chicken and dog). Somewhat surprisingly, AUGUSTUS +T is not very much behind the methods that use human mRNAs besides other evidence. For example, AUGUSTUS +T has a gene-level accuracy of only 10 percentage points lower than JIGSAW at approximately the same gene-level specificity, however JIGSAW used human mRNA, human and non-human RefSeqs, the UCSC KnownGenes, the ENSEMBL predictions, PhastCons and the predictions of six other gene finders, which themselves include further evidence. We have to note that the human genome as a test case is somewhat special. The non-human RefSeq mRNA annotations could have been influenced by the knowledge of the annotators of the human genes, so that possibly a circularity effect might lead to an overestimation of accuracy. On the other hand, non-human mammalian target genomes would benefit from the availability of the well-annotated human informant genome. Table S3 shows the accuracy when all mRNAs from Genbank of the five informant species are used and not only the RefSeq mRNAs.

This new version of AUGUSTUS has been applied to annotate the genes in *Galdieria sulphuraria* (manuscript in preparation). In this genome project, we incorporate extensive cDNA sequence data sequenced with 454 technology besides protein-level conservation [protein hints not discussed here, see Stanke *et al.* (2006b)]. We find that the method presented here also works well when using the shorter ESTs from 454 sequencing.

4 DISCUSSION

In this study, we used the human genome to evaluate our program for several reasons. Besides the direct interest in the human genome, we chose it because it is relatively well annotated for a complex genome and many gene prediction tools are adjusted and have previously been evaluated using standard reference sets. However, the many newly sequenced genomes, in which the genes need to be indentified, have less available transcript evidence, in particular much fewer full length mRNA sequences than human. For many species, even a little native EST data exists. Therefore, restricting the gene prediction in human to the usage of different evidence sources can be seen as an estimate of the performance in other species where direct gene prediction assessment is more difficult.

De novo gene finding is useful mostly for finding genes or parts of genes for which insufficient evidence exists from transcript or protein data. The *de novo* method presented here is modular and simple. Nevertheless, it approximately achieves the accuracy of N-SCAN, the previously most accurate *de novo* gene finder for human. Further, PhastCons is not a specific method for finding conserved protein coding regions. We only included it to supplement ExoniPhy as it sometimes misses conserved exons. We expect that further improvements can be achieved by using a more inclusive prediction of evolutionary conserved exons.

Using only ESTs and genomic conservation, AUGUSTUS is significantly more accurate than N-SCAN/EST. However, each of the two programs can be seen as an extension of a GHMM-based comparative gene finder, AUGUSTUS +X,p,RA and N-SCAN, respectively, which have similar performances. The main difference between the two methods of incorporating EST alignments is that AUGUSTUS interprets these alignments as information about *intervals* and N-SCAN/EST interprets them as information about individual *positions*. N-SCAN/EST reduces the information of the EST alignments to a so-called *ESTseq*, a set of likely intronic bases and a disjoint set of likely exonic bases. When predicting a gene structure, it rewards or penalizes the classification of individual bases as exonic or intronic based on the *ESTseq*. So, a predicted intron having overlap with many intronic bases from the *ESTseq* may be rewarded although it actually contradicts the splicing suggested by EST alignments. In contrast to the method proposed here, the N-SCAN/EST approach also loses the information about possibly complicated alternative spicing and nesting of genes by projecting the EST alignments to a single *ESTseq*. In light of these differences, we attribute the advantage of AUGUSTUS X,p,E over N-SCAN/EST to a more careful method of evidence integration.

5 CONCLUSION

We conclude that our gene annotation pipeline is particularly strong in two major settings: First, when the major source of gene evidence for the genome is ESTs and, second, when one or more well-annotated informant genomes are available that are related closely enough to show synteny.

We showed that despite the quality issues of ESTs and their fragmented nature, EST-supported gene finding can be much more accurate than previously shown in mammals. Further, the accuracy, mainly the specificity, can be increased using genomic conservation in addition to the EST evidence. We expect that ESTs will play an even more important role for the annotation of future genomes because cheaper new sequencing methods allow for economically obtaining a good coverage of the transcriptome.

The indirect use of alien cDNA by mapping it syntenically to the target genome can produce almost as accurate gene models as methods that use native transcripts, even when exceptionally many and full-length native mRNAs are available as in the case of human. This method is very well-suited to annotate, e.g. the mammalian genomes, making use of the relatively good human, mouse and rat annoations, and the fact that usually only little native transcript data is available. In upcoming genome projects this setting will also become more relevant as for most finished genomes there will be suitable informant genomes that have been annotated before. Also, with AUGUSTUS native and alien transcripts can be combined and complement each other to achieve even higher accuracy.

ACKNOWLEDGEMENTS

This work was supported by a fellowship within the Postdoc Program of the German Academic Exchange Service (DAAD) to M.S. This work was partially supported by Federal Ministry of Research and Education (BMBF) project 'MediGRID' (BMBF 01AK803G). This project has been funded in whole or in part with Federal Funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

We thank Michael Brent and Jeltje van Baren for giving us the N-SCAN/EST+PASA predictions, Adam Siepel for providing the ExoniPhy and PhastCons tracks, and the anonymous referees for their helpful suggestions to improve the manuscript.

Conflict of Interest: none declared.

REFERENCES

Allen,J.E. and Salzberg,S.L. (2006) A phylogenetic generalized hidden Markov model for predicting alternatively spliced exons. *AMB*, **1**, 14.

Brejová,B. (2005) Evidence combination in hidden Markov models for gene prediction. PhD Thesis. University of Waterloo, Canada.

Brejová,B. *et al*. (2005) ExonHunter: a comprehensive approach to gene finding. *Bioinformatics*, **21** (Suppl. 1), i57–i65.

Cawley,S.L. and Pachter,L. (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19** (Suppl. 2), ii36–ii41.

Curwen,V. *et al*. (2004) The Ensembl Automatic Gene Annotation System. *Genome Res*., **14**, 942–950.

Djebali,S. *et al*. (2006) Exogean: a framework for annotating protein-coding genes in eukaryotic genomic DNA. *BMC Genome Biol*., **7** (Suppl. 1), S7.1–10.

Florea,L. *et al*. (2005) Gene and alternative splicing annotation with AIR. *Genome Res*., **15**, 54–66.

Foissac,S. and Schiex,T. (2005) Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics*, **6**, 25. Available at: http://www.biomedcentral.com/1471-2105/6/25.

Gross,S.S. and Brent,M.R. (2005) Using multiple alignments to improve gene prediction. In *Proceedings of RECOMB 2005*. Springer, Berlin, pp. 374–388.

Guigó,R. *et al*. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *BMC Genome Biol*., **7** (Suppl. 1), S2.1–31.

Haas,B.J. *et al*. (2003) Improving the Arabidopsis genome annotation using maximal transcipt alignment assemblies. *Nucleic Acids Res*., **31**, 5654–5666.

Harrow,J. *et al*. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol*., **7** (Suppl. 1), S4.1–9.

Kent,W.J. (2002) BLAT–The BLAST-Like Alignment Tool. *Genome Res*., **12**, 656–664.

Kent,W.J. *et al*. (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *PNAS*, **100**, 11484–11489.

Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, S1–S9.

Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI, pp. 179–186.

Kuhn,R.M. *et al*. (2006) The UCSC genome browser database: update 2007. *Nucl. Acids Res*., **35**, D668–D673.

Lomsadze,A. *et al*. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucl. Acids Res*., **33**, 6494–6506.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–80.

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

Meyer,I.M. and Durbin,R. (2004) Gene structure conservation aids similarity based gene prediction. *Nucl. Acids Res*., **32**, 776–783.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res*., **35** (Suppl. 1), D61–65.

Schwartz,S. *et al*. (2003) Human-Mouse Alignments with BLASTZ. *Genome Res*., **13**, 103–107.

Siepel,A. and Haussler,D. (2004) Computational identification of evolutionarily conserved exons. *Proceedings of RECOMB 2004*, p. 177–186.

Siepel,A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.

Siepel,A. *et al*. (2007) Targeted discovery of novel human exons by comparative genomics. *Genome Res*., **17**, 1763–1773.

Stanke,M. and Waack,S. (2003) Gene prediction with a hidden markov model and new intron submodel. *Bioinformatics*, **19** (Suppl. 2), ii215–ii225.

Stanke,M. *et al*. (2006a) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*., **34**, W435–W439.

Stanke,M. *et al*. (2006b) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA supported gene and transcripts annotation. *BMC Genome Biol.*, **7** (Suppl. 1), S12.

van Baren,M.J. and Brent,M.R. (2006) Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res*., **16**, 678–685.

Wei,C. and Brent,M.R. (2006) Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics*, **7**, 327.

Zhu,J. *et al*. (2007) Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Computational Biol*., **3**, e247. Available at: http://dx.doi.org/10.1371/journal.pcbi.0030247.