Published online 5 March 2009 Nucleic Acids Research, 2009, Vol. 37, No. 7 e52
doi:10.1093/nar/gkp052

Finding genes in *Schistosoma japonicum*: annotating novel genomes with help of extrinsic evidence

Broňa Brejová[1], Tomáš Vinař[2], Yangyi Chen[3], Shengyue Wang[3], Guoping Zhao[3,4], Daniel G. Brown[5], Ming Li[5] and Yan Zhou[3,6,*]

[1]Department of Computer Science, [2]Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics, Comenius University, Mlynska Dolina, 84248 Bratislava, Slovakia, [3]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, [4]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, [5]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada and [6]School of Life Sciences, Fudan University, Shanghai 200433, China

Received September 13, 2008; Revised and Accepted January 15, 2009

ABSTRACT

We have developed a novel method for estimating the parameters of hidden Markov models for gene finding in newly sequenced species. Our approach does not rely on curated training data sets, but instead uses extrinsic evidence (including paired-end ditags that have not been used in gene finding previously) and iterative training. This new method is particularly suitable for annotation of species with large evolutionary distance to the closest annotated species. We have used our approach to produce an initial annotation of more than 16 000 genes in the newly sequenced *Schistosoma japonicum* draft genome. We established the high quality of our predictions by comparison to full-length cDNAs (withdrawn from the extrinsic evidence) and to CEGMA core genes. We also evaluated the effectiveness of the new training procedure on *Caenorhabditis elegans* genome. ExonHunter and the newest parametric files for *S. japonicum* genome are available for download at www.bioinformatics.uwaterloo.ca/downloads/exonhunter

INTRODUCTION

Schistosoma japonicum is one of three human parasitic organisms from the phylum Platyhelminthes (flatworms) that cause schistosomiasis. This disease is responsible for 15 000–20 000 deaths every year and is endemic in 76 countries of the world (1). In this article, we present the novel methods that were used to predict the protein-coding genes in the nearly 400 Mb draft genome sequence of *S. japonicum* (*S. japonicum* Genome Consortium unpublished data).

We concentrate on the problem of estimating the parameters of a hidden Markov model (HMM)-based gene finder. Ordinarily, we would require a high-quality training set of several hundred genes to reliably estimate the necessary parameters, such as the nucleotide composition in coding and noncoding regions, the length distributions of introns, exons and intergenic regions and a characterization of splice signals. In the initial stages of genome projects, however, such a training set is rarely available. Another approach is to use parameters estimated from genes of a related species; this is common practice in vertebrate genomics. However, this is not feasible in the case of *S. japonicum*. Its evolutionary distance from currently annotated genomes precludes this approach (Figure 1), as its closest annotated relative is at a distance of more than 500 Myr to the most recent common ancestor.

We propose an iterative training procedure that first uses gene predictions in our new species, found by a gene finder trained with parameters from a distantly related, but well annotated organism (in our case, *Caenorhabditis elegans*) as a starting point. These initial predictions, which are often of poor quality, are then used to retrain the parameters of the gene finder. The retraining step is repeated several times, to build an incrementally improving training set.

A similar approach has been first examined on the simple *ab initio* gene finder SNAP, by Korf (2), and later re-evaluated in a more practical setting using GeneMark.HMM by Lomsadze *et al.* (3). Here, we

*To whom correspondence should be addressed. Tel: +86 21 50804801; Fax: +86 21 50801922; Email: zhouy@chgc.sh.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

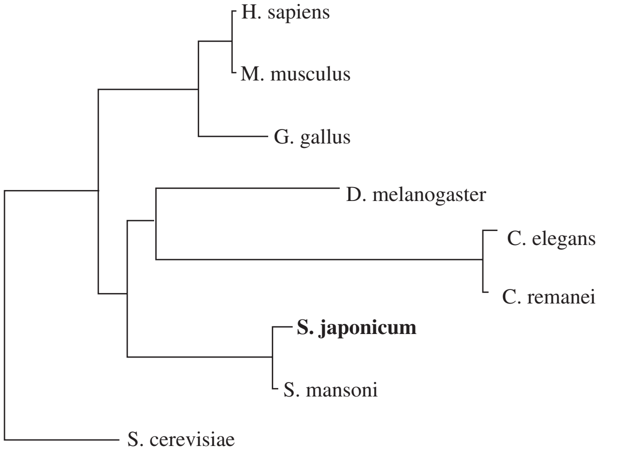

Figure 1. Evolutionary distance of *S. japonicum* from well–annotated species. The phylogeny was derived by maximum likelihood from a multiple alignment of small ribosomal subunit RNAs (9) using PHYML (10) and MUSCLE (11).

introduce a new method that incorporates extrinsic evidence to this iterative process. Extrinsic evidence, such as expressed sequenced tags (ESTs) and protein databases, is known to significantly improve gene predictions when integrated with an *ab initio* gene finder (4). Our work extends the use of extrinsic evidence from gene prediction to the training of gene finders. In particular, we use the additional evidence to locate fragments of predictions that are likely to be of high quality, and we use these fragments for training in the next iteration of our iterative training.

To validate our approach, we train and test our gene finder ExonHunter (5) on the *C. elegans* genome, where we can easily evaluate our methods with a reliable testing set. We compare the *ab initio* performance of ExonHunter to SNAP (2) and GeneMark.HMM (3) and find that iterative training with filtering by extrinsic evidence significantly improves the quality of *ab initio* predictions. We also note that use of the extrinsic evidence in both training and gene prediction leads to further significant improvements in prediction accuracy.

Using iterative training with extrinsic evidence, we estimated parameters for ExonHunter in newly sequenced *S. japonicum*, and produced initial protein-coding gene predictions for further analysis. We also used a set of full-length cDNAs, withdrawn from the training set and from our extrinsic evidence sets, to evaluate accuracy of our predictions on *S. japonicum*.

One of the most difficult problems in gene finding is to correctly identify the boundaries of individual genes (4). To address this problem, we used a new source of extrinsic evidence, Paired-End diTags (PETs). A single PET consists of two 14–21 bp tags from both the 5′ to 3′ ends of a transcript. Ng *et al.* (6) introduced a high-throughput genome-wide method for sequencing PETs by modification of the SAGE approach. We used a set of PETs available for the *S. japonicum* genome to improve gene predictions by integrating them into ExonHunter's existing framework for extrinsic evidence.

Another approach to gene finding in a novel genome was recently introduced by Parra *et al.* (7). The authors identified 458 genes that are well-conserved among a wide variety of species, and they propose a pipeline, CEGMA, that can identify these genes in newly sequenced genomes by profile alignment methods. This core set of genes can then be used to train a general purpose gene finder. Here, we use the CEGMA pipeline to predict the set of core genes in *S. japonicum* and compare the set with genes predicted by ExonHunter. The two methods are orthogonal and can be both used with varying success in newly sequenced species for producing high-quality initial gene predictions, but we show that in *S. japonicum* genome, ExonHunter recovers more core genes with higher accuracy than CEGMA.

In fact, the CEGMA core genes have been used as a starting point within the annotation pipeline MAKER (8) (developed in parallel with our approach), which combines the *ab initio* gene finder SNAP (2) with cDNA and EST alignments. Within this pipeline, a similar idea to our iterative training is used to refine the SNAP HMM by retraining on a small subset of predicted genes showing high concordance with ESTs. Here, we provide a more general methodology that can be used with a much larger variety of extrinsic evidence, and also in scenarios where sparse evidence supports mostly short sections of individual genes (as is the case in *S. japonicum* genome). We also demonstrate that weaker cross-species evidence can often be enough to train a reliable gene finder using our methods.

MATERIALS AND METHODS

ExonHunter and extrinsic evidence

ExonHunter combines a generalized HMM for gene finding with extrinsic evidence from a variety of sources, such as EST databases, databases of known proteins and repeat databases (5). Briefly, extrinsic evidence is summarized in the form of a *super advisor* that assigns a probability distribution over the set of sequence labels Σ to each position in the sequence; the super advisor is meant to be independent of local sequence features. The set of labels Σ represents annotation features, such as exon, intron or intergenic region; ExonHunter uses a richer label set that also includes reading frame, strand and the signals at exon boundaries. Since different labels from Σ occur in gene annotations with different genome frequencies (for example, coding regions comprise only a small portion of vertebrate genomes), we divide the probability of each label ℓ in the super advisor by the prior probability of the label ℓ to obtain a super advisor score s_ℓ. If this score is >1 (super advisor assigned a probability to label ℓ higher than its prior), the HMM probabilities are manipulated so that ExonHunter is more likely to predict the label ℓ for the current DNA base. On the other hand, if the score is <1, ExonHunter will attempt to avoid that label at that position. Figure 2 shows an example of several sources of extrinsic evidence and the resulting ExonHunter gene prediction.

The position-by-position probabilities of the super advisor are obtained by a combination of multiple *advisors*, each representing an individual source of extrinsic evidence. A given advisor may not have enough information to specify a full probability distribution over all labels in Σ. Thus, at each sequence position, the advisor only gives a probability distribution over a partition of the set of sequence labels Σ. For example, if a particular position of the sequence is spanned by a protein–DNA sequence alignment, an advisor corresponding to the protein database will assign probability P to the coding label ℓ with the matching coding frame and probability $1 - P$ to the set containing all other labels. The probability P is estimated from training data and depends on the score of the protein match. If no protein match covers the sequence position, the advisor will make the *vacuous* statement that the complete set of labels Σ has probability 1, which does not contain any information.

To combine advisors into the super advisor, we have previously used a method based on quadratic programming (5). In this article, we use a simpler method based on linear combination of all advisor statements. At each position, advisors that issued vacuous statements are removed, and the predictions of the remaining advisors are simplified by distributing the probability assigned to each set S of labels to individual labels $s \in S$ proportionally to the prior probabilities of those labels. The super advisor probability distribution is then estimated by a linear combination of these simplified advisor predictions. We have previously tested this method and found its performance satisfactory and robust to changes of parameters of individual advisors (12). In the final step, the super advisor scores are combined with the probabilistic distribution defined by the HMM (5) and the modified most probable annotation of the sequence is predicted by a modified Viterbi algorithm.

Extrinsic evidence used for *C. elegans*. The following databases have been used as a source of the external evidence for *C. elegans* training and gene predictions: RepeatMasker RepBase for *C. elegans* genome, 30 919 *C. elegans* ESTs from TIGR gene index (v.9, 22 September 2004), 20 292 *C. remanei* ESTs downloaded from Genbank (2 February 2007), proteins for selected species from SWISSPROT release 52.0 (3000 *C. elegans* proteins, 15 803 human proteins and 2473 *Drosophila melanogaster* proteins) and the Pfam database (v.11.0).

The genomic sequences were screened for the RepBase repeats with RepeatMasker (www.repeatmasker.org) and the output was processed as described by Brejova *et al.* (5). ESTs were first prescreened by wublast (blast.wustl.edu), and then aligned to the genomic sequences by sim4 (13). Proteins were aligned to the genomic sequences by blastx (14). Finally, we have aligned protein domain profiles from the Pfam database by rpsblast.

S. japonicum repeat library. We have used RepeatScout (15) to identify families of repetitive sequences that occur in *S. japonicum* genome. We required that each repeat occurs in the genomic sequence at least 12 times; this threshold was determined by examining the repeat overlap with our library of ESTs. The resulting repeat library contains more than 600 different sequences. We used RepeatMasker to align the repeat sequences from the library to the *S. japonicum* genome. Repeats found in this way were then used as an advisor to lower probability of coding regions overlapping repetitive sequences.

Extrinsic evidence used for *S. japonicum*. In addition to the repeat library described above, we have used the following sources of extrinsic evidence for *S. japonicum* training and gene predictions: 86 139 unclustered *S. japonicum* ESTs sequenced for *S. japonicum* genome project, 40 683 ESTs from TIGR *S. mansoni* gene index (release 6.0, 19 June 2006) and 126 640 unclustered ESTs sequenced for the *S. mansoni* genome project, 21 772 SWISSPROT proteins from human, *C. elegans*, yeast and schistosomas. We aligned these with the *S. japonicum* genome in the same way as for the *C. elegans* data described above. In addition, we have used PETs described below as a source of extrinsic evidence.

Paired-End diTags and their mapping

Paired-End diTags are sequences with a mean length of 35 bp, each containing the 5′ and 3′ signatures of a full-length transcript (6,16) and a collection of 307 056 distinct PETs were obtained (unpublished data).

We mapped the PETs to the genome by using wublast. We filtered the wublast results to include only matches where the 5′ and 3′ signature had a total of at least 28 bases aligned to the same genomic contig, spanning a region between 150 and 100 000 bases long. This resulted in approximately 110 000 matches. More than 80% of these PETs mapped to a unique location in the genome.

We clustered the PET-spanned regions according to their genomic overlaps. Many PETs support the same transcript or alternative transcripts of the same gene. In each cluster, we chose only a single representative supported by the most PETs with the end coordinates within 20 bp of each other. Moreover, we required that each locus is supported by at least two PETs having a unique match in the genome, or at least six PETs mapping to multiple locations. The resulting set contains 4746 loci, out of which 4336 are supported by uniquely matching PETs. In some ExonHunter predictions, we have used a simpler mapping pipeline, which did not determine global uniqueness of PETs. However, such variations in PET mapping did not significantly affect ExonHunter predictions.

Full-length cDNA *S. japonicum* testing set

Of the full-length cDNAs 17 186 (unpublished data) were sequenced for the purpose of the *S. japonicum* genome project. We have mapped all cDNAs longer than 300 bp to the genome using blat (17), requiring 95% coverage at 95% sequence identity. This has resulted in 23 305 alignments of 10 989 cDNAs. We have further checked these alignments for canonical splice sites, and in each alignment we have located the longest open reading frame with length of at least 300 nt. Finally, out of the 5584 resulting alignments (not necessarily nonredundant), we have selected 951 nonoverlapping transcripts that cover the largest amount of the DNA sequence. These sequences represent a high-quality testing set of transcripts. In this article, we have decided not to evaluate the effects of alternative splicing, thus our testing set contains only a single transcript per gene.

The above pipeline for mapping the testing set is designed with a set of sequence analysis tools that is distinct from those that we used to map ESTs to the genome. In this way, we are looking to avoid potential bias in evaluation.

RESULTS

Iterative training with extrinsic evidence

A gene finder can be trained by an iterative process, where instead of the training set one uses the predictions produced with a previous iteration of the parameters. This method is also called Viterbi training, and while it is not guaranteed to converge to the set of parameters maximizing the likelihood of the data, it has been shown to produce reasonable performance in gene finding (2,3). We have developed a new iterative technique that takes into account extrinsic evidence, such as protein homology or ESTs. We explain the method in the context of our gene finder ExonHunter (6), though we note that it is applicable more generally.

Gene predictions supported by extrinsic evidence are typically much more accurate and therefore they are more suitable as a training set for a gene finder than unsupported *ab initio* predictions. To train our gene finder, we use an iterative method similar to Viterbi training. However, in each iteration, we use the super advisor (see Materials and methods section) to identify fragments of gene predictions that are well supported by extrinsic evidence, and we only use these as a training set for the next iteration.

To identify supported fragments, we classify each position as *supported*, *in conflict* or *unknown*. Labels at supported positions agree with available extrinsic evidence, which is reflected in their high super advisor probability. In particular, a position is supported, if its predicted label has the highest score among all labels at that position, and the score is above certain threshold (we use 1.01). Position is in conflict if some other label is supported at that position. We classify a predicted exon or intron as supported, if at least half of its bases are supported and no base is in conflict. An exon or intron is in conflict if at least one base is in conflict. For training purposes, we use this classification to select the *fragments* of our predictions that are best supported by extrinsic evidence. In particular, we select all supported exons and combine them into longer chains, if they are connected by introns that are not in conflict. Each such chain is used as a separate, potentially incomplete transcript in the filtered training set. An example of the supported fragments is shown in Figure 2.

We use only the supported fragments to estimate all parameters of the HMM, except for intergenic lengths whose distribution is estimated from the complete unfiltered predictions. Unfiltered predictions are also used to train prior distribution of super advisor and all other parameters for super advisor framework. Using the fragments would lead to underestimation of prior probabilities of coding regions and introns.

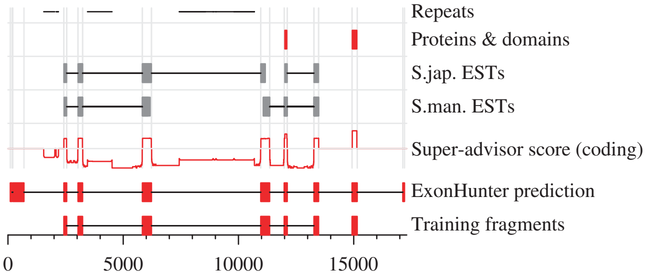

Figure 2. Selection of supported gene fragments. ExonHunter integrates several sources of extrinsic evidence (such as sequence repeats, known proteins, ESTs and PETs). The figure shows an example of alignments of individual sources to genomic sequence at a particular locus. This information is combined into the super advisor (super advisor score for coding regions is shown). The same super advisor scores are used to aid in gene prediction, and to identify supported fragments of the gene for training.

To validate our approach, we tested several variants of our iterative training approach on well annotated *C. elegans* genome. We used 38 Mb portion of the genome for training and additional 6 Mb for testing. As a baseline, we trained ExonHunter on curated WormBase genes overlapping the training sequence. Then we used ExonHunter to predict genes in the testing sequence *ab initio* (using only sequence repeat information), with *cross-species* evidence (using *C. remanei* ESTs and Swissprot proteins from species other than *C. elegans* s extrinsic evidence) and *all evidence* (using all of the above, as well as *C. elegans* ESTs and proteins). We measured exon and gene sensitivity by comparing these predictions to curated WormBase genes overlapping the testing set (see Figure 3 and Supplementary Table S1). We can see that the accuracy of gene finding improves as we increase the amount of extrinsic evidence. Ideally, we would be able to achieve similar levels of sensitivity by our iterative method without using the curated genes for training.

To verify this claim, we started with ExonHunter parameters as trained on the human genome (using RefSeq genes) and improved the parameters iteratively, measuring sensitivity at each iteration (iteration 0 represents the predictions with the original human genome parameters). We evaluated several variants of iterative training, and the results are shown in Figure 3.

First, we have considered simple Viterbi training. We used ExonHunter in *ab initio* mode and used unfiltered predictions to estimate parameters for the next iteration. This strategy does not reach accuracy comparable to *ab initio* ExonHunter trained on the curated genes, even after three iterations. However, we can see that this is not an issue specific to the ExonHunter design, since similar or worse accuracy is achieved by Viterbi training of SNAP (2) and GeneMark.HMM (3) (Table 1).

On the other hand, our new iterative training strategy allows us to use all available evidence, both in predictions and in selection of supported fragments for further iterations. With this change, we achieve almost the same

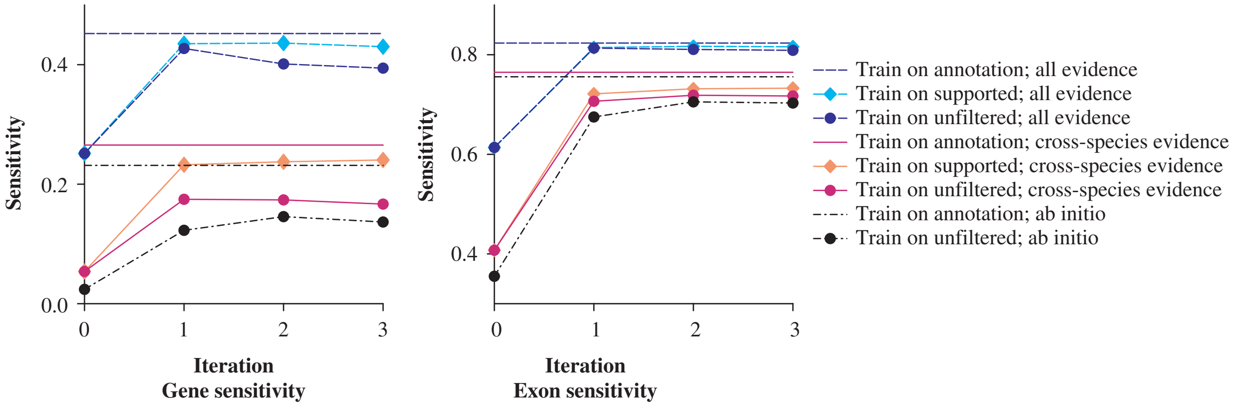

Figure 3. Evaluation of iterative training on *C. elegans* (gene and exon sensitivity). Each line in the plots is annotated with the method used for training and the level of extrinsic evidence used for both training and testing. Using supported gene fragments helps iterative training to achieve the performance close to the training on a curated training set. The filtering step is important especially when working with weak extrinsic evidence. Specificity comparison leads to the same conclusion, see Supplementary Data. In all experiments on *C. elegans*, ExonHunter under-predicted the number of genes, as can be seen by comparison of gene sensitivities and specificities in Supplementary Table S1.

Table 1. Accuracy of *ab initio* gene finders using different supervised and unsupervised training strategies on a portion of the *C. elegans* genome

	Gene (%)		Exon (%)		Int. exon (%)		Nucleotide (%)	
	sn	sp	sn	sp	sn	sp	sn	sp
Iterative training without filtering								
ExonHunter (3 iter.)	17	23	72	72	84	75	94	93
SNAP (3 iter.)	8	5	63	45	74	50	89	81
GeneMark HMM ES	20	23	70	73	81	78	94	91
Iterative training, filtering by extrinsic evidence								
ExonHunter (3 iter.)	24	28	73	72	84	75	94	91
Traditional training with training set								
ExonHunter	23	30	76	74	87	76	95	92
SNAP	21	17	70	70	78	75	92	92

Among methods using unsupervised iterative training, ExonHunter and GeneMark perform at similar levels of sensitivity and specificity. Using extrinsic evidence and filtering in iterative training (the new feature introduced in this article) improves performance of *ab initio* gene finding mainly on gene level; the performance approaches that of supervised training with curated training sets. Sensitivity (sn) is a percentage of annotated features correctly predicted by each method. Specificity (sp) is a percentage of predictions that are true positives. Internal coding exon (int. exon) sensitivity and specificity excludes first and last coding exons which are typically annotated less accurately.

accuracy as with curated training data set, already after the first iteration (Figure 3). This improvement is not explained simply by added evidence in prediction step on the testing set; we have also improved underlying HMM parameters, since *ab initio* version of ExonHunter with these parameters shows significant improvement compared to Viterbi-trained ExonHunter (at the gene level), SNAP and GeneMark.HMM (Table 1). Moreover, adding evidence only in the prediction step to the model trained by the *ab initio* iterative training also yields lower prediction accuracy, especially on the gene level (Supplementary Table S3).

Extrinsic evidence for *C. elegans* includes a rich set of *C. elegans* proteins and ESTs, which greatly simplifies the task of gene prediction. In a newly sequenced genome, we do not expect to have such strong sources of extrinsic evidence. Therefore, we have also evaluated the iterative training using only cross-species evidence, excluding *C. elegans* ESTs and proteins. Compared to *ab initio* predictions, ExonHunter trained on curated genes performs only slightly better. However, even weak evidence leads to considerable improvement in iterative training, particularly at the gene level. In this setting, we also see the greatest advantage of using the supported fragments for training compared to an unfiltered set.

PETs as extrinsic evidence

One of the hardest problems in gene finding is to correctly identify the boundaries of genes. Computationally, the effort to resolve this problem has concentrated on sophisticated models for identification of translation start sites (18), 5′ untranslated region model improvements (19) and attempts at accurate identification of transcriptions start site (20,21). Here, we present a different approach by employing PETs (7), experimental evidence that can be obtained at high throughput and at low cost.

PETs are pairs of 14–21 bp sequence tags extracted from both ends of an mRNA transcript. To achieve high-throughput identification of PETs, they are ligated to form larger concatemers. Sequencing of each concatamer

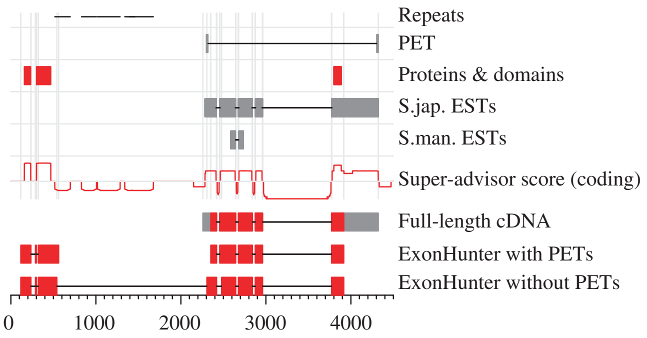

Figure 4. Influence of PETs on ExonHunter predictions. The PET mapped to the genome sequence correctly identifies the extent of the transcript supported by full-length cDNA (the transcript includes untranslated regions shown as shaded areas). The prediction of ExonHunter without PETs incorrectly identifies start of the gene and adds two spurious exons to the transcript. Using PETs not only helps to identify the start site, but also corrects the reading frame of the first exon and acceptor site of the second exon.

clone can identify 30 or more PETs. For the purpose of the *S. japonicum* project, 688 210 PETs were sequenced, yielding 307 056 distinct tags with average multiplicity of 2.2. Mapping of the PETs to the current *S. japonicum* assembly yielded 4735 clusters of PETs, with average span of 10 193 bp (see Materials and methods section).

A typical relationship of a PET and a gene structure is shown in Figure 4. We have compared our set of 4735 PETs to a set of 957 genes supported by full-length cDNA transcripts (see Materials and methods section). Out of 812 PETs that intersect at least one full-length cDNA transcript, 622 (76%) match both transcript ends within 100 bp tolerance and 512 (63%) within stricter 30 bp tolerance. The PETs cover 65% of the genes in the full-length cDNA set (54% with 30 bp tolerance), although we expect that the overall genome coverage will be lower, since both cDNA sequencing and PET sequencing is biased toward genes with high expression.

Mapping of PETs to the genome provides two kinds of extrinsic information for gene finding. First, the tags map to gene boundaries and so they should be adjacent to an intergenic region. In ExonHunter, we use this information by elevating the probability of the intergenic region label within short regions outside both tags in a mapped PET (extending for 150 bp starting at 10 bp outside of identified tags). Additionally, a typical PET spans the full extent of a transcript, and so the region between two paired tags should contain exactly one gene. To include this information in ExonHunter, we would need to increase the probability of all labels except intergenic over the whole extent of the gene. In preliminary experiments, this led to undesirable artifacts in predicted gene structures, so we have not used this feature in this article. Thus, our current use of PETs discourages ExonHunter to extend a gene beyond boundaries determined by the PET, but ExonHunter can still predict multiple genes or no gene at all in that region.

The use of PETs slightly improves exon sensitivity and specificity of ExonHunter (roughly by 1%); however, it increases the number of completely correctly predicted genes in our full-length cDNA set from 325 to 361 (roughly by 4%). Figure 4 shows how an ExonHunter prediction can change by using the PETs.

Another simple use of PETs is to cut out the portion of the genome between the two tags and running a gene finder on such sequence fragments. ExonHunter, as well as many other HMM-based gene finders, can be easily modified so that it predicts at most one gene in a given sequence. We have used this mode to predict genes in some of the PETs that did not overlap any predictions.

S. japonicum gene predictions and evaluation

To predict protein-coding genes in the newly sequenced *S. japonicum* genome, we have applied two iterations of the training strategy described above, starting from ExonHunter parameter set estimated from the *C. elegans* curated gene set. We have used ESTs from *S. japonicum* and *S. mansoni*, homology to Swissprot proteins (human, *C. elegans*, *S. cerevisiae* and *Schistosomas*), Pfam domains, PETs and a custom sequence repeat library (see Materials and methods section) as the extrinsic evidence.

In each of the two iterations, we have run ExonHunter on the draft of *S. japonicum* genomic sequence. The final predictions were produced by running ExonHunter with all extrinsic evidence and applying a simple postprocessing, where we filtered out predictions with significantly conflicting external evidence and those that were very short. After the filtering, we were left with 16 687 predicted genes.

To evaluate the quality of our predictions, we have assembled a set of 951 nonoverlapping full-length cDNAs with 3860 coding exons, sequenced as part of the *S. japonicum* genome project (see Materials and methods section). These cDNAs were excluded from our extrinsic evidence so they did not affect our predictions directly. However, the cDNAs may partially overlap independently obtained ESTs used as extrinsic evidence.

The predicted coding regions have a median length of 651 bp and median number of exons 3. This is comparable to the cDNA-testing set with a median coding length 657 bp and median number of exons 4. Table 2 shows a comparison of the predictions overlapping the testing set and the gene structures based on the full-length cDNAs. ExonHunter predicts 75% of the exons in our cDNA set correctly. The fraction grows to 85% when we consider only internal exons. Start and stop codons are harder to predict than splice sites, and at the same time their annotation in our cDNA set is less reliable. More than a third of the cDNA genes are predicted completely correctly. We have also evaluated specificity by considering all predicted exons that are located at most 100 bp beyond a testing gene boundary. Roughly 66% of them coincide exactly with a cDNA exon. As we see in the table, the prediction accuracy has increased substantially after the first iteration of the training, but no further accuracy gain was achieved by the second iteration.

The genes predicted by ExonHunter were further curated, resulting in a set of 12 657 genes (unpublished data) that were used for further *S. japonicum* genome analysis. Table 2 shows that the curation process further increased specificity of the predictions.

Table 2. Accuracy of ExonHunter iterative training on *S. japonicum*

Iteration	Gene (%)		Exon (%)		Int. exon (%)		Nucleotide (%)	
	sn	sp	sn	sp	sn	sp	sn	sp
0	31	28	71	56	80	56	92	80
1	38	35	75	66	85	68	93	85
2	38	35	75	66	85	68	93	85
Curated	40	40	73	70	82	72	91	86

We compare our gene predictions to a set of 951 full-length cDNAs. The accuracy improves in the first training iteration and remains level afterwards. Selected ExonHunter predictions were included in a human-curated annotation which has somewhat higher accuracy than raw ExonHunter results. Sensitivity (sn) is the percentage of annotated features correctly predicted by each method. Specificity (sp) is the percentage of predictions that are true positives. Specificity is computed by considering only gene predictions that overlap the loci covered by the cDNA sequences and their flanking regions (see the text for more details).

Note that the genes in the full-length cDNA set are more likely to be represented in the EST set due to the same sampling bias toward highly expressed genes, and therefore the prediction accuracy on this small evaluation set is likely higher than the overall genome average. Indeed, 78% of gene predictions overlapping the full-length cDNAs have at least half of their coding positions supported by the evidence, whereas only 44% of all predicted genes achieve this level of support.

One advantage of ExonHunter is its ability to seamlessly combine evidence from multiple sources. Figure 5 shows an example of a gene supported by ESTs on both ends, whereas the middle part is supported by *S. mansoni* ESTs and a Swissprot match. A PET delineating gene boundaries was also used as evidence.

We have also compared the ExonHunter predictions to the core gene set identified by CEGMA (7). CEGMA is designed to find genes in a newly sequenced organism based on profiles of 458 eukaryotic conserved orthologous groups (KOGs) (22) that are well-conserved over a variety of six eukaryotic species (human, *Drosophila*, *Arabidopsis*, *C. elegans*, *S. cerevisiae* and *S. pombe*). These genes can serve as a training set and as an indication of genome completeness. By running CEGMA on *S. japonicum* genome, we have obtained 264 core gene predictions. The CEGMA core genes overlap 70 cDNAs in our testing set, giving us an opportunity to compare the quality of the CEGMA core genes and ExonHunter predictions (all of these cDNAs also overlap an ExonHunter prediction). ExonHunter achieves higher prediction accuracy than CEGMA (Table 3), although due to the small size of the testing set, the improvement is not statistically significant (gene level sensitivity, $P = 0.08$, sign test). We have observed that ExonHunter predictions seem to be more accurate on transcript ends. Figure 5 shows an example where ExonHunter predicted extension is supported by protein evidence (i.e. is likely coding), and the resulting protein matches the CEGMA protein profile better than the corresponding CEGMA core gene.

Only 264 CEGMA core genes (out of possible 458 KOGs) were found in *S. japonicum*. This number is lower than in the other species tested by Parra *et al.* (7); the smallest number of core genes (∼300) was predicted in the protozoan parasite *Toxoplasma gondii*, while in each of the other three eukaryotic species tested (*Anopheles gambiae*, *Ciona intestinalis* and *Chlamydomonas reinhardtii*), CEGMA predicted more than 400 core genes. The low number of genes found in *S. japonicum* can be either due to fragmented sequence assembly, large divergence from the species represented in the profiles, atypical gene loss or errors in the CEGMA predictions. Indeed, Parra *et al.* (7) ascribe the low number of core genes in *T. gondii* to the large divergence from the other reference genomes.

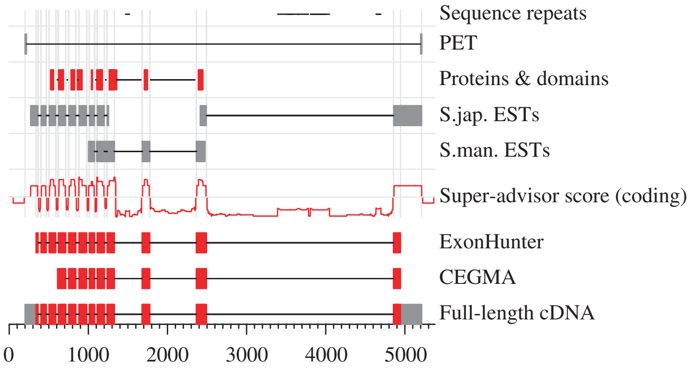

Figure 5. CEGMA and ExonHunter predictions of vacuolar proton pump subunit C homolog. ExonHunter combines evidence from *S. japonicum* and *S. mansoni* ESTs, as well as SWISSPROT protein to predict the gene structure. The ExonHunter prediction of this gene is of better quality than the core gene predicted by the CEGMA pipeline.

By running CEGMA's final filtering step on the ExonHunter predictions instead of the predictions produced by the CEGMA pipeline, we were able to predict core genes for 59 additional KOGs. Therefore, we conclude that the CEGMA pipeline misses some genes due to gene prediction errors. In contrast, we have found only six genes predicted by CEGMA that either did not occur among ExonHunter predictions, or the corresponding ExonHunter prediction did not match the KOG profile sufficiently to pass the filter. The total number of core genes discovered either by ExonHunter or the CEGMA pipeline is 323 (which is an increase of 22% compared to the CEGMA core genes).

Table 3. Accuracy of CEGMA and ExonHunter compared to a set of 70 full-length cDNAs

Iteration	Gene (%)		Exon (%)		Int. exon (%)		Nucleotide (%)	
	sn	sp	sn	sp	sn	sp	sn	sp
ExonHunter	64	61	88	84	92	87	99	96
CEGMA	53	52	80	82	81	83	94	95

Sensitivity (sn) is a percentage of annotated features correctly predicted by each method. Specificity (sp) is a percentage of predictions that are true positives. CEGMA core gene predictions only overlap 70 full-length cDNAs; we are comparing the quality of ExonHunter and CEGMA predictions on this set. To evaluate specificity of ExonHunter, we have only considered gene predictions that overlap the loci covered by these cDNA sequences and their flanking regions (see the text for more details).

DISCUSSION

With the arrival of rapid low-cost sequencing technologies (23), we face the imminent challenge of analyzing increased number of previously uncharted genomes. Producing high-quality gene annotations of both well-characterized and novel protein-coding genes is one of the first steps required for such analysis. Since the characteristics of protein-coding genes vary between species, most gene-finding techniques require a high-quality curated training set of several hundred genes, before the parameters of a gene finder can be trained for a particular genome.

In this article, we proposed a new technique that does not rely on such large training sets. Instead, we use extrinsic evidence, together with an iterative training method, to replace the training set. This way, we can produce high-quality gene annotations by an automated process in early stages of the genome project. Using our techniques, we have produced initial gene annotations for *S. japonicum* genome, and we found those predictions to be of high quality by comparison to full-length cDNA sequences produced in later stages of the project.

Earlier work (2,3) concentrated on similar iterative methods for training *ab initio* gene finders. However, gene finders incorporating additional evidence, including ESTs and protein databases, have been previously shown to produce superior gene predictions (4). By introducing this additional evidence to both the iterative training providing more reliable training sets, and the final gene prediction step, we have achieved significantly higher accuracy of the final predictions.

Using large amounts of data is essential in training more complex models of splice sites and other sequence elements (24). For example, instead of iterative training, one option is to train *S. japonicum* gene finders on a portion of the 951 full-length cDNA set now available. Even though this set is already larger than many training sets used in gene finding, it only contains 3860 coding exons, whereas we were able to identify approximately 33 000 supported exons in each iteration of our training method. Use of this larger number of exons for training may allow for far more richly parametrized models.

While ESTs and protein databases are commonly used as a source of information in gene prediction, PETs have not been used in gene prediction yet. PETs can be obtained by a low-cost high-throughput protocol and can provide essential information about the boundaries of individual transcripts and alternative starts and ends of transcribed regions. Even with our simple approach of incorporating PETs as a source of evidence, we achieved modest improvements in accuracy. Perhaps an even greater improvement could be achieved by methods for evidence combination that could use PETs more fully. PETs indicate only the extent of a transcript and need to be combined with other sources of information to specify the full exon–intron structure of the gene. At the same time, they introduce long-range dependencies over the whole extent of a gene, whereas gene finders typically consider evidence independently for every position, as in ExonHunter (5), or for every exon and intron, as in Augustus (25) and Jigsaw (26). Therefore, development of methods capable of using incomplete information spanning whole genes or even larger regions remains a challenge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Yijun Ruan and the Genome Institute of Singapore for the kindness of providing PET data, Gang Lu and Huajun Zhen for the pre-processing of PET data, and Prof. Yijun Ruan for discussion on the genomic mapping of PET data.

FUNDING

China National High Tech Program (863) (project numbers 2006AA02Z335 to Y.Z., 2008AA02Z313 to M.L.); Shanghai Municipal Commission of Science and Technology; Chinese Academy of Sciences; National Key Project for Basic Research (973); National Natural Science Foundation of China; Canada Research Chair program (to M.L.); National Science and Engineering Research Council of Canada (OGP0046506 to M.L.); Commission of the European Communities (224885 to T.V. and 231025 to B.B.). Funding for open access charge: 863-2006AA02Z335.

Conflict of interest statement. None declared.

REFERENCES

1. World Health Organization Expert Committee (1993) The control of schistosomiasis. *Technical report* 830. WHO technical report series.
2. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
3. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y.O. and Borodovsky,M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, **33**, 6494–6496.
4. Guigo,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B., Birney,E. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7(Suppl. 1)**, 1–31.
5. Brejova,B., Brown,D.G., Li,M. and Vinar,T. (2005) ExonHunter: a comprehensive approach to gene finding. *Bioinformatics*, **21(Suppl. 1)**, i57–i65.
6. Ng,P., Wei,C.-L., Sung,W.-K., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A., Wong,C.H. *et al.* (2005) Gene identication signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, **2**, 105–111.
7. Parra,G., Bradnam,K. and Korf,I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
8. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Sanchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
9. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam,S.A., McGarrell,D.M., Garrity,G.M. and Tiedje,J.M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33**, D294–D296.
10. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
11. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
12. Brejova,B. (2005) *Evidence Combination in Hidden Markov Models for Gene Prediction*. Ph.D. Thesis, University of Waterloo.
13. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
15. Price,A.L., Jones,N.C. and Pevzner,P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21(Suppl. 1)**, i351–i358.
16. Chiu,K.P., Wong,C.H., Chen,Q., Ariyaratne,P., Ooi,H.S., Wei,C.L., Sung,W.K. and Ruan,Y. (2006) PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics*, **7**, 390.
17. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
18. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19(Suppl. 2)**, ii215–ii215.
19. Brown,R.H., Gross,S.S. and Brent,M.R. (2005) Begin at the beginning: predicting genes with 5′ UTRs. *Genome Res.*, **15**, 742–747.
20. Sonnenburg,S., Zien,A. and Ratsch,G. (2006) ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, **22**, e472–e480.
21. Ohler,U. (2006) Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.*, **34**, 5943–5950.
22. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
23. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
24. Vinar,T. (2005). *Enhancements to Hidden Markov Models for Gene Finding and Other Biological Applications*. Ph.D. Thesis, University of Waterloo.
25. Stanke,M., Schoffmann,O., Morgenstern,B. and Waack,S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
26. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.