Genome Annotation in Plants and Fungi: EuGène as a Model Platform

Sylvain Foissac[1,2], Jérôme Gouzy[3], Stephane Rombauts[4,5], Catherine Mathé[2], Joëlle Amselem[6,7], Lieven Sterck[4,5], Yves Van de Peer[*,4,5], Pierre Rouzé[8] and Thomas Schiex[*,9]

[1]Centre de Regulació Genòmica, PRBB, Aiguader 88, 08003 Barcelona Catalunya, Spain; [2]Unité Mixte de Recherche 5546, Centre National de la Recherche Scientifique-Université Paul Sabatier-Toulouse III, Pôle de Biotechnologies Végétales, 31326 Auzeville, France; [3]Laboratoire Interactions Plantes Micro-organismes UMR441/2594, INRA/CNRS, F-31320 Castanet Tolosan, France; [4]Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium; [5]Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium; [6]Unité de Recherche Génomique-Info, INRA, 91000 Evry, France; [7]Unité Mixte de Recherche BIOGER, INRA AgroParisTech, 78000 Versailles, France; [8]Laboratoire Associé de l'INRA France, Ghent University, 9052 Ghent, Belgium; [9]Unité de Biométrie et Intelligence Artificielle, UR 875, INRA, F-31320, Castanet Tolosan, France

Abstract: In this era of whole genome sequencing, reliable genome annotations (identification of functional regions) are the cornerstones for many subsequent analyses. Not only is careful annotation important for studying the gene and gene family content of a genome and its host, but also for wide-scale transcriptome and proteome analyses attempting to describe a certain biological process or to get a global picture of a cell's behavior. Although the number of sequenced genomes is increasing thanks to the application of new technologies, genome-wide analyses will critically depend on the quality of the genome annotations. However, the annotation process is more complicated in the plant field than in the animal field because of the limited funding that leads to much fewer experimental data and less annotation expertise. This situation calls for highly automated annotation platforms that can make the best use of all available data, experimental or not. We discuss how the gene prediction (the process of predicting protein gene structures in genomic sequences) research field increasingly shifts from methods that typically exploited one or two types of data to more integrative approaches that simultaneously deal with various experimental, statistical, or other *in silico* evidence. We illustrate the importance of integrative approaches for producing high-quality automatic annotations of genomes of plants and algae as well as of fungi that live in close association with plants using the platform EuGène as an example.

Keywords: Genome annotation, fungi, plants, gene finding.

1. INTRODUCTION

An accurate annotation of the genes in the genome of an organism is essential for downstream bioinformatics analyses and for the design of genome-wide biological assays. The rapid expansion of the rate of genome sequencing has led to increased reliance on fully automated methods. The shift from semi-automated annotation, requiring a large panel of expert annotators, to fully automated systems has probably occurred first in the plant field because of the more limited funding, but is now becoming an important issue in all genome annotation projects.

A key issue to find protein-coding genes is the ability to handle the diverse evidence available for gene prediction. Following successful application in gene prediction [1], the mathematical model of Generalized Hidden Markov Models (GHMMs) has been used in most accurate gene finders. HMMs define a probabilistic framework for modeling random sequences composed of different homogeneous regions. Initially developed for speech analysis, it is widely used in sequence analysis and general algorithms for training (estimation of parameters in the model) and segmenting (detecting homogeneous regions) have been developed [2]. The efficiency of these algorithms fundamentally relies on independence assumptions that are not always acceptable. Some GHMM-based systems are specialized in dealing with specific types of evidence, such as the conservation of exons. In two different related species, functional regions such as exons tend to be more conserved than other regions. This difference in conservation is exploited in software such as TwinScan [3] and NSCAN [4].

In practice, curators routinely incorporate very diverse data, such as available predictions, expressed sequence tag (EST)-spliced alignments (a DNA/DNA alignment that tries to open gaps on consensus splice sites corresponding to possible introns), protein similarities, comparative sequence, or other species-specific evidence. However, such information is very difficult to capture in GHMMs, because they rely on independence and locality assumptions (see also [5]) and, therefore, cannot be simply decomposed in independent contributions or incorporate long-range effects.

To integrate such knowledge, the simplest way is to use a pipeline that simultaneously gathers all the evidence and different predictions through dedicated genome browsers, allowing researchers to evaluate the different evidence tracks. Some recent systems, such as JIGSAW [6], try to consolidate a prediction by integration of different sources of

*Address correspondence to these authors at the Unité de Biométrie et Intelligence Artificielle, Unité de Recherche 875, INRA, 24 Chemin de Borde Rouge, F-31326, Auzeville, France; E-mail: Thomas.Schiex@toulouse.inra.fr and Bioinformatics and Evolutionary Genomics, VIB, Ghent University, Technologiepark 297, B-9052 Gent, Belgium;
E-mail: Yves.vandepeer@psb.ugent.be

1574-8936/08 $55.00+.00

evidence. Nevertheless, a more rational design would directly merge all the available evidence into a single gene prediction system that should also incorporate new evidence in a easy and rapid manner.

2. EUGÈNE AS AN INTEGRATIVE GENE PREDICTION PLATFORM

The need for an automatic prediction procedure that matches the quality of the current structural annotation of the yeast genome was the initial and central motivation for the development of EuGène [7], a highly integrative protein-coding gene prediction platform for eukaryotes. The problem of gene prediction can be defined as a segmentation problem, in which the aim is to subdivide the genomic DNA into genomic regions of different types. The simplest protein gene model discriminates coding exons (the coding part of any exon) and the remainder of the sequence, considered as noncoding. In eukaryotic organisms, most often, noncoding regions are further divided into introns that separate coding exons and other noncoding regions, allowing clustered coding exons to be properly bordered for each individual gene. Each of these regions is usually delimited by specific patterns or motifs, such as splice sites, translation initiation sites, etc.

Instead of *a priori* matching a given mathematical model to the gene finding problem, so as to be classified as a purely statistical or computational problem, EuGène was designed after observation and analysis of an expert annotator at work. The aim of this analysis was

(1) to collect all the sources of evidence exploited for gene prediction;

(2) for each piece of evidence, to identify the type of information provided by the data useful for gene prediction;

(3) to recognize the satisfactory set of properties, independently of the process to be used by each annotator to make a prediction.

The typical list of evidences for expert annotation can be divided easily into three categories:

(1) statistical evidence, or more generally, *in silico* evidence provided by dedicated mathematical models with the purpose of catching the specific properties of gene components. The simplest models provide local information that a region seems to be *coding* for a protein (Markov models [8]) or that a given possible splice site looks functional or not, based on a classifier (e.g. Window Array Models [9] or Support Vector Machines [10]). These models underlie all existing *ab initio* gene finders. Beyond this local information, complete gene predictions built by existing gene finders are often used for expert prediction as well.

(2) Similarity with sequences of documented molecules that are the gene expression products is often the most accepted evidence for detecting gene structure. Typically, different types of similarities are used in the gene prediction process. First, high similarities with known proteins usually clearly indicate that the sequence considered is *coding*. Second, similarities with expressed sequences (ESTs or cDNAs built from mRNAs from the same or a closely related organism) strongly suggest that the genomic sequence is potentially *transcribed* and conserved after mRNA processing (although it might not code for proteins). These two types of similarity are illustrated in Fig. (**1**) Finally, similarities with repeated regions and, especially, documented transposable elements, are often used to identify spurious *in silico* predictions.

(3) Similarity of the DNA sequence itself with that of other genomes can result from a selective pressure throughout evolution, suggesting a functional role for the conserved region. For example, it might be used to identify *coding* regions. Such a region is

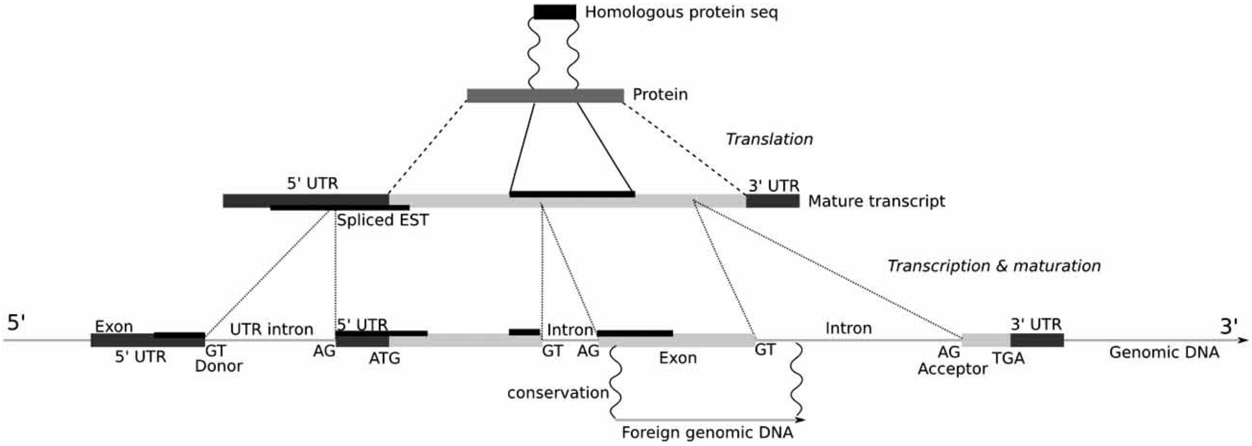

Fig. (1). A general view of the different levels of access to biological information on protein genes. Here a gene is shown on the genomic DNA, with exons and introns going through (bottom-up) transcription, maturation and translation processes (in italic). Experimental evidence may appear at the mature transcript level: a spliced EST (partial matured RNA sequence), shown as a black box on the left of the figure, can be aligned on the genomic DNA allowing the detection of an intron (UTR intron here). Similarly, the existence of a similarity (shown by snake-like relations, top of the figure) at the protein level can be mapped back, although usually less precisely, to give a hint that the corresponding genomic sequence is translated. Also, possible conservation with a foreign genomic DNA sequence is a fuzzy indication that the corresponding region may be functional (typically a coding exon – which is part of the translated region of the mRNA).

represented in Fig. (**1**). These conserved regions usually do not exactly match coding regions. The conservation may extend to introns or may represent another type of functional region (noncoding gene for example).

One should note that each source of evidence bears information from different times of the gene expression and protein synthesis process. As Fig. (**1**) shows, while protein similarity is an indication that the sequence can reach the translation step and is probably part of a coding exon, mRNA similarity is an indication that the sequence has undergone transcription and maturation. In the latter case, such a sequence cannot be intergenic nor intronic, but could code or not for a 5' or 3' untranslated regions (UTRs) of a gene (Fig. **2**).

Precise identification of similarities between sequences is much more difficult in eukaryotic organisms because of the fragmented nature of their genes than in prokaryotes. However, reliable tools are now available for so-called spliced alignments between DNA or protein sequences and genomic DNA with high efficiency and precision, including the identification of small exons (e.g. GeneSeqer [11] or GenomeThreader [12]).

All these sources of evidence are obviously interdependent in complex ways that are extremely difficult to capture or model. However, GHMM, the most frequent mathematical model used for gene prediction [1], assumes that the evidence and the given sequence are independent. To address this intrinsic limitation, GHMM-based gene finders often require before final release some dedicated –and not always documented-tuning that is not compatible with an automatic annotation system.

Furthermore, a lesson learned over the last decade, is that sources of evidence evolve rapidly over time because of the development of new technology either on the experimental or on the *in silico* side. Therefore, it is crucial for a gene finder to easily integrate new arbitrary sources of evidence. Although this requirement is not always feasible for computational reasons, the underlying methods and the software architecture should be designed as to make this integration as simple as possible. Here again, the flexibility of the widely used GHMM has shown some limitations in the past. For instance, to incorporate additional types of evidence, such as similarities with protein sequences or with other genomic sequences into the Genscan system [1], specific methods have been developed to produce different software tools, such as Genomescan [13] and Twinscan [3], respectively. Research in comparative genomics even dictated a considerable change of the inherent model, as in NSCAN [14]. Ideally, a powerful annotation system would be expected to integrate different kinds of evidence in a convenient and generic fashion. Regarding the above-mentioned examples, it is worth noting that these software tools are independent and that none of the GHMM-based gene finders yet has the possibility to exploit simultaneously more than one of these types of evidence (AUGUSTUS+, one of the most advanced software in this category, only uses the most reliable information at each position). More fundamentally, essential changes in the assumed properties of possible gene structures might be needed (as illustrated by the radical change in the importance of alternative splicing over the last years).

Based on all available evidence, EuGène uses a three-step approach for the gene prediction itself:

(1) all possible segmentations of the genomic DNA in regions that define a consistent gene structure (with respect to the current biological knowledge for a given organism) are represented concisely in a weighted directed acyclic graph. The conciseness is very important because the number of such structures grows exponentially with the sequence length. The representation must remain flexible and should ideally only grow linearly with the sequence length.

(2) A score must be assigned to choose among all these consistent gene structures based on the available evidence. To build a parametric scoring system, the quality of prediction on the documented data set itself is optimized, differing from the usual maximum likelihood optimization often

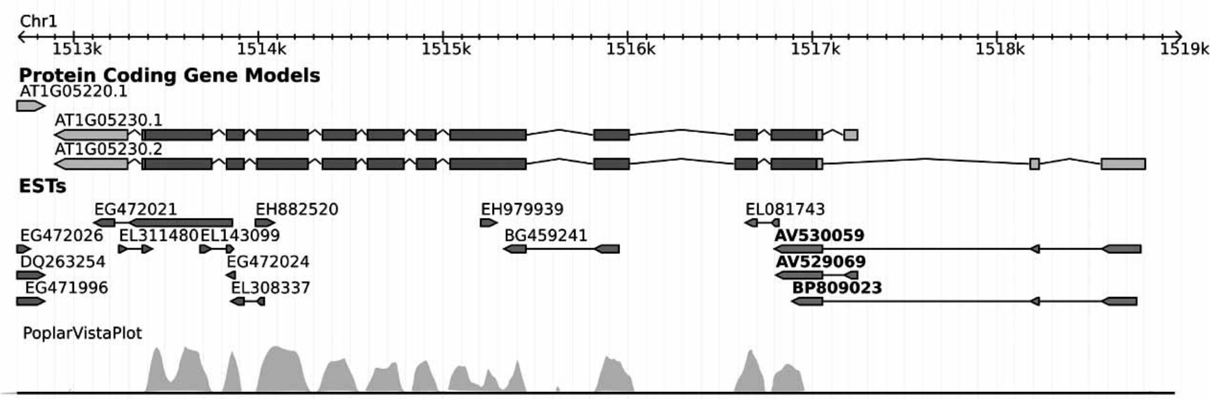

Fig. (2). Real example of the experimental (nonstatistical) evidence (EST similarities and conservations) usually available for gene structure prediction. Beyond the noisy nature of conservation itself (here between *Arabidopsis thaliana* and poplar), the experimental evidence may appear internally inconsistent because ESTs (AV529069 and AV530059/BP809023) are in contraction on the existence on a transcribed region between position 1517k and 1518k. These two ESTs represent a likely alternative splicing event. Dealing with such a situation is difficult for most integrative gene finders. Moreover, the gene here contains introns in both the 5' UTR and the 3' UTR that are rarely modeled in practice.

used in GHMM-based gene finders (frequently with additional manual parameter tweaking).

(3) Once the scoring system is defined, given a genomic sequence and the set of all available evidence, a gene structure with optimal score is identified among the exponential number of possible predictions. Again, to obtain efficient algorithms, the conciseness of the gene structure representation is crucial and produces an optimal prediction. Alternative splicing detection usually means that several gene predictions might be necessary to explain inconsistencies in expression evidences.

2.1. Modeling Gene Structures

Different sources of evidence can provide information for gene prediction. GHMMs and protein similarities are informative, whether a region codes or not, while ESTs, mRNAs, or tiling arrays capture information after RNA processing and reveal which parts of the genome are transcribed. These various points of views calls for a gene model that effectively represents all possible situations with respect to *transcription*, *splicing*, and *translation*. Therefore, the gene model of an integrative gene finder should be able to represent and discriminate coding exons, introns that separate coding exons, and the remainder of the genome. In addition and because of directly informative sources of evidence at this level, a sophisticated gene finder must classify regions that are *transcribed* (or not) and those that are *spliced* out (or not) during maturation, requiring that 5' and 3' UTRs as well as UTR introns (introns that separate two exons that contain UTRs) are taken into account. The gene model used in EuGène effectively discriminates between all these regions. The ability of the gene model to deal with all these different types of sequence is important because otherwise, the existence, for example, of an intron in an UTR based on a cDNA-spliced alignment would falsely predict two coding exons bordering the detected intron.

An extremely simplified graph used for representing gene structures in EuGène (Fig. **3**) is defined for a specific genomic sequence and omits UTR introns and the negative strand. Its primary property is that any path from left to right is a consistent gene structure. It is essentially the unrolled version of the automata that underlies a GHMM. The actual gene model includes a total of over 40 different possible region types, because intron phases, coding frames, and strandness are taken into account and first, internal, last, and single coding exons are handled separately (Fig. **4**). Beyond splice sites, translation start and stop, transcription start and stop, information about possible frameshifts and length distributions can also be included in EuGène.

As a consequence of the increasing importance of comparative genomics for gene prediction, the number of region types represented alongside protein-encoding regions must obviously be augmented. An integrative gene finder should be able to denote regions, such as regulation sites (not transcribed, but usually conserved) or noncoding RNAs (transcribed, at least conserved at the secondary structure level, but without UTR). To our knowledge, only one gene finder, SLAM [15], explicitly considers conserved noncoding regions, but essentially to avoid false positive predictions.

2.2. Learning A Scoring System

To select the most likely predicted gene structure given the available evidence, a score should be assigned to every possible structure. If every edge in the graph receives a score, the score of a gene structure could be defined as the sum of the local scores associated to the edges of the path. Each of these local scores (or votes) themselves will be defined by all the sources of evidence. For modularity and ease of implementation, each source of evidence used in a given application of EuGène is represented by a plug-in (a dynamically loaded software component). On a nucleotide-per-nucleotide basis, each plug-in scores horizontal edges, associated to the fact that, in a given type of segment or diagonal edges, a nucleotide represents occurrences of signals (all sorts of functional sites).

These local votes should favor or disfavor structures that are compatible or incompatible with the evidence, respectively. As an illustration, let us consider the integration of a splice site detection program, such as SpliceMachine [10], into a EuGène plug-in. SpliceMachine produces a score for every possible canonical AG/GT acceptor and donor. This

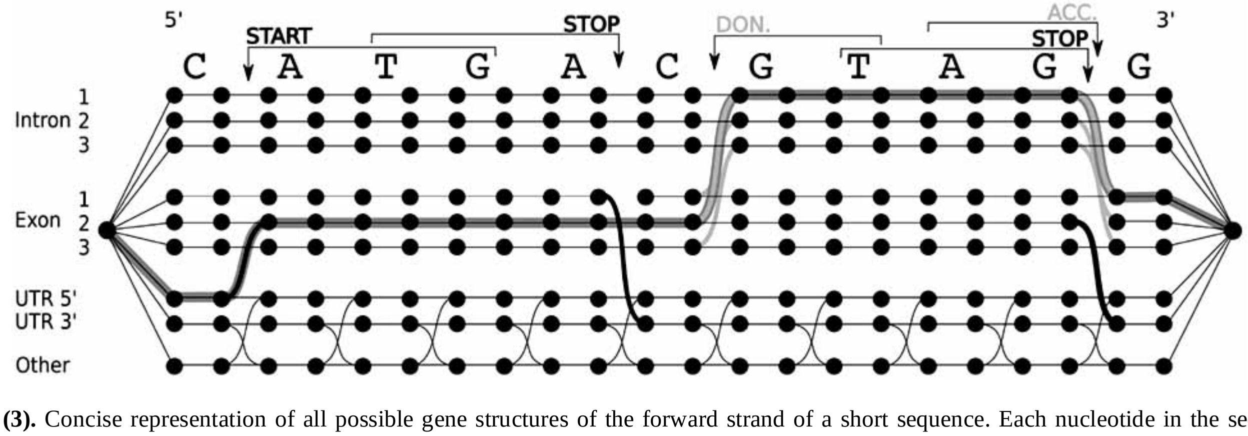

Fig. (3). Concise representation of all possible gene structures of the forward strand of a short sequence. Each nucleotide in the sequence (top) can be either intergenic, in a 5' or 3' UTR, in a coding exon, or at the three possible phases of an intron (corresponding to the splicing of a codon in the first, second, or third position), represented by the different tracks. Occurrence of so-called signals (ATG for translation start, GT/AG typically for donors and acceptors...) allow a segment to be started or ended by changing tracks; for example, an in-phase ATG at the second position (phase 2) allowing the shift from the 5' UTR to the coding exon in phase 2, is represented by a diagonal edge between positions 1 and 2 that connect the two corresponding tracks. Each left-to-right path without back steps corresponds to a possible gene structure.

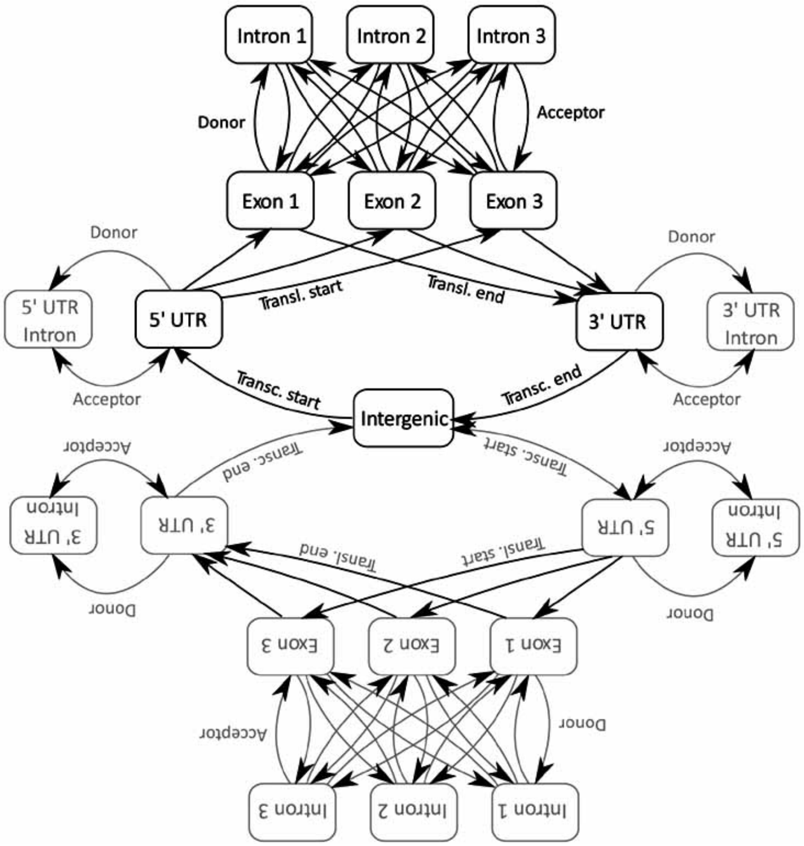

Fig. (4). Complete view of all the possible states in EuGène. The graph of Fig. **2** only illustrated the dark states, but ignored the UTR introns and the opposite Strand.

score contributes to the weight of the diagonal edges that represent possible splicing at every occurrence of a canonical AG/GT splice site position. Similarly, an order *k* 3-periodic GHMM [8] computes the probability $P(X_i|X_{i-1},...,X_{i-k})$ that the nucleotide X_i at position *i* appears in an exon, given the *k* previous nucleotides and this for every possible phase. The opposite of the logarithm of these probabilities defines different votes that can contribute to the score of the horizontal edges associated with the nucleotide X_i in each coding phase. In practice, these votes can have different semantics and are often not independent. For all these reasons, the direct combination of all these scores would probably not lead to an effective gene prediction tool. Therefore, each vote *s* produced by a given source of evidence *p* is rescaled with a parametric function before being combined. The simplest rescaling functions are linear functions $f_p(s) = a_p.s$ with one parameter, or affine functions $f_p(s) = a_p s + b_p$ with two parameters, but more complex functions could be used as well.

In gene finding, the quality of a prediction tool is often assessed empirically on a set of sequences annotated by experts. The sensitivity (*Sn*) of a gene prediction tool is defined as the percentage of known (annotated) functional elements that are predicted as such, whereas its specificity (*Sp*) is the percentage of predicted elements that are known as functional. A perfect prediction has 100% sensitivity and specificity. However, the two measures are not independent: higher sensitivity is often obtained at the price of lower specificity; they can be computed at the nucleotide level (considering coding nucleotides as functional), at the exon level (an exon prediction must match the annotation exactly to be correct), or at the gene level (all coding exons of a gene predicted exactly). These definitions ignore alternative mechanisms (such as splicing), but are nevertheless very informative. The parameter estimation in EuGène attempts at maximizing the geometric mean of the gene and exon level sensitivity and specificity defined as $\sqrt[4]{Sne\ .\ Spe\ .\ Sng\ .\ Spg}$. The geometric mean has been chosen because a high geometric mean implies that none of its term is itself close to zero.

Optimization of this mean is very difficult because the optimized function is neither differentiable nor even continuous. Therefore, meta-heuristics (with a genetic algorithm) have been combined with a heuristic search process based on coordinate descent.

Overall, the scoring system used in EuGène is very flexible: it easily captures GHMMs (with log-probabilities and without rescaling) and it deals with nonindependent sources. For example, when the same source of information is used twice (completely correlated sources), parameter re-estimation simply divides the optimal rescaling parameters among the two sources and the prediction quality remains unchanged. This instance also shows that the solved optimization problem is also degenerated in the sense that usually more than one single optimal set of parameters will be found.

Because scores or votes (also called potentials) are intimately related with probabilities (for example, through the Boltzman distribution), the above scoring system might be interpreted also probabilistically. As a consequence, because EuGène is a purely discriminative system and handles arbitrarily nonindependent observation, it is also related to Semi-Markov Conditional Random Fields [16].

2.3. Building A Prediction

Once all parameters are estimated, building a prediction is just a matter of finding an optimal path in a directed acyclic graph (Fig. **3**), which can be done easily with the linear time, linear space Bellman-Kalaba algorithm [17], or the simple variant, the GHMM Viterbi algorithm [2]. In EuGène, a slightly more sophisticated version allows scores to be integrated depending on the lengths of the predicted segments, like in GHMMs. The algorithm used in EuGène is still linear in space and time in the sequence length, similarly as the gene prediction system AUGUSTUS [18]. For actual gene predictions as described in Section 4, the length distributions used are uninformative distributions that just forbid short introns. This is probably what makes possible the prediction of many new short genes compared to the *Arabidopsis thaliana* annotation of TAIR [19,20].

Compared to the original version described in [7], EuGène has been extended to represent new region types (including UTR and UTR introns), has been made capable of incorporating new sources of evidence such as conservation [21], new site prediction software (for splice sites, translation starts...), of producing GFF3 validated output... Beyond the possibility to integrate various information sources, EuGène is equipped with a large Perl tool kit for pre- and post-processing, for training and for direct generation of TIGR_XML outputs. Finally, EuGène can include information on transcript and protein similarities, and recently it has been extended to predict alternatively spliced variants, based on inconsistent EST data [22].

3. A REVIEW OF EXISTING INTEGRATIVE SYSTEMS

Building a consensus annotation from available evidence is not new and several gene finding systems utilize comparable approaches. Only truly integrative programs will be considered that are directly capable of integrating different types of information. Several gene finders will be excluded that "just" incorporate statistical (*ab initio*) and conservation information of related organisms (software such as TwinScan or NSCAN)

To better understand how integrative gene finders relate, they can be classified along few lines. First, the inherent uncertainty of information sources can be managed in different ways and can be based on probabilities (or by scores that can be related to probabilities with Boltzman distribution) or on a nonpurely numerical representation. For instance, the ENSEMBL pipeline [23] tries to incorporate first what is considered as the most reliable information (protein similarities). Therefore, the reliability granted to information sources is essentially captured by the pipeline structure itself, which is difficult to change or optimize.

Most other existing integrative gene finders incorporate information in a global manner, by quantifying the relative reliabilities of all the available information sources with probabilities or scores, e.g. EuGène [7], but also GAZE [24], AUGUSTUS+ [25], JIGSAW [6], and Conrad [5]. These systems go through a training phase to capture the reliability of information sources and then through a prediction phase that produces an optimal prediction. In almost all existing tools, training is performed on a set of curated sequences, each containing one or more known genes together with associated evidence. The set must be as representative of the genome gene contents as possible, avoiding overrepresentation of one gene family or of simple (intronless) genes and it must also be of sufficient size, typically hundreds of genes. For an in-depth analysis of the influence of the training set size on prediction performances in GHMM-based gene finders, see [26] and [27].

However, the training phase and the type of information that can be captured during the training vary a lot. As in many machine learning or statistical problems in bioinformatics, integrative gene finding faces two-dimensionally organized data: one (horizontal), the sequence itself and the other (vertical), the list of evidence available at each position. For a given source of evidence, horizontally short or long-range dependencies might occur, while at a given position, arbitrary sources of evidence might correlate vertically.

AUGUSTUS+ uses a generalized GHMM with additional hints [25] that are assumed essentially independent both horizontally and vertically. This model, without extrinsic information, is already capable of producing *ab initio* predictions. The training procedure that quantifies weights associated to each source of information is a maximum likelihood approach with a training set to assess independently the reliability of each source. When several sources of evidence are available, only the most reliable is used. Despite its apparent simplicity, AUGUSTUS+ gave excellent results in the recent ENCODE evaluation [28].

The approach of EuGène differs from that of AUGUSTUS+ by the utilization of the maximum of empirical success as the only criteria for parameter estimation. Although it is much more difficult to optimize, this criterion avoids, at least to some extent, the assumption of vertical independence between sources of evidence although one or two parameters per source is obviously not enough to capture complex correlations between sources. The criteria used in the Semi-Markov Conditional Random Field-based gene finder Conrad [5] are similar to those in EuGène, despite different optimization algorithms for parameter estimation. Naturally, these gene finders are extremely sensitive to the training set used. Because curated gene sets have usually a lot of attached evidence (making curation possible), a direct training will underestimate the importance of purely statistical evidence when strong experimental evidence seems to be always available and extremely informative. Specific tricks have been applied, such as removal of part of the available evidence for training (Conrad) or incremental estimation of parameters, such as adding each type of evidence after another (EuGène).

From the point of view of capturing vertical correlations, the most advanced software is JIGSAW that uses decision trees, but, like GAZE, is not a proper gene predictor, because it cannot produce a list of possible coding regions alone. Therefore, JIGSAW can better be considered as an integration tool, because it relies on existing gene predictions that can be integrated together with other evidence. It is best adapted to situations in which a lot of information and different good quality *ab initio* predictions are already available, which is rather unusual for plants and fungi.

Finally, there are usually only minor differences in the underlying principles of the prediction phase that always exploits the linear structure of genomic sequences in dedicated dynamic programming algorithms. AUGUSTUS+, EuGène, GAZE, Conrad, and JIGSAW, all allow optimal predictions to be found based on an additive score that incorporates local scores and a function describing the length of the regions predicted. The latter allows capturing the actual length distribution of exons, introns and other region types.

4. EXAMPLES OF PLANT AND FUNGAL GENOME ANNOTATION EFFORTS WITH EUGÈNE

Applications of EuGène to three different organisms, namely *Arabidopsis thaliana*, *Medicago truncatula*, and the plant-associated organism *Botrytis cinerea* (fungal pathogen) will be described below. When EuGène is run for a specific organism, it usually relies on a different set of sources of evidence. A typical basic recipe includes GHMMs and splice and translation site predictors that define a purely *ab initio* application of EuGène. When EST, mRNA, and protein similarities are included, the advantages of each databank should be considered. Furthermore, gene duplication is rampant in plant genomes, leading to many gene families with many members [29]. Therefore, the knowledge of one gene might be useful to predict divergent, but homologous, gene occurrences. Similarly, when related organisms have already been sequenced, conservation with these organisms can be utilized. In addition, existing reliable *ab initio* gene finders might also contribute to the overall prediction quality. Furthermore, the specific resources of each organism are often worth exploiting, emphasizing the need for a flexible integrative annotation system.

4.1. Eugène Used to Annotate *Arabidopsis thaliana*

EuGène was first trained on the *A. thaliana* genome. To evaluate the effect of the increase in information on the prediction quality, different increasingly large sets of plug-ins have been tried on this genome. The first variant of EuGène used NetStart [30] for ATG prediction, both NetGene2 [31] and SplicePredictor [32] for splice site prediction, and an internal interpolated Markov model [33] plug-in for coding and noncoding regions. For the evaluation, the Araset dataset [34] was used. The Markov models had been trained previously on another data set defined as the union of Araclean [35] and a set of more than 1,000 manually curated genes, based on full-length cDNA, designated PlantGene, after removal of all sequences already present in Araset. The rescaling parameters were optimized on the 144 genes of Araclean only, defining a first *ab initio* variant of EuGène.

A more sophisticated version integrated similarities with *A. thaliana* ESTs and mRNAs extracted from dbEST (April 2002; 123,160 sequences) and PlantGene. Spliced alignments were built with sim4 [36]. Because EST data are very noisy, the plug-in for the analysis of the EST similarities in EuGène uses a pre-filtering process that first sorts EST alignments according to an increasing order of detected introns and throws away any EST alignment incompatible with previous alignments in the above order. The EST plug-in is based on alignments to vote on splice sites and on regions transcribed (UTR and coding exons) on matching regions and spliced out for gaps. As previously, the strength of the votes has been optimized on Araclean that, together with the *ab initio* plug-ins mentioned above, defined the EST variant of EuGène.

Alternatively, protein similarities are very useful. A simple plug-in analyzes the output of NCBI-BlastX [37] against a given protein databank. This plug-in determines whether a region codes in the frame of the match and is intronic on the border of successive matches. The strength of the decision or vote is obtained by linear rescaling of the similarity score. This plug-in was applied to SwissProt, PIR, and SP-TrEMBL. After optimization, the rescaling parameters of SP-Trembl had a value of 0 and was therefore removed. The final protein variant of EuGène used just a filtered version of SwissProt with all *Arabidopsis* proteins removed for a fair evaluation, whereas the FULL variant of EuGène utilized all *ab initio* plug-ins, EST, mRNA, and protein similarities.

Table **1** lists the different gene finders tested. An evaluation of the sensitivity and specificity at the gene and exon level on Araset for these different gene prediction software and the different variants of EuGène is presented in (Table **2**). These tests show that the *ab initio* variant of EuGène is already quite good, and only slightly worse than FgenesH at the exon level. The integrative variant with EST and proteins correctly predicts 77% of the 168 genes in the data set. The gene level specificity of 73% is probably underestimated, given the number of strong protein hits in regions initially annotated as intergenic in Araset.

Before this evaluation, EuGène had been applied on the complete genome and the predictions were used to design the genome wide CATMA microarray [38], into which additional information was integrated. First, repeat-containing regions (as detected by RepeatMasker [39]) were exploited. The dedicated plug-in votes against the fact the matching regions are coding to avoid spurious predictions. Second, the Riken Institute produced especially extended full-length cDNAs sequenced at both extremities that produced EST couples [40]. These sequences were aligned to the genome and EuGène was forced to predict a single gene in the corresponding region by a dedicated plug-in that votes against "intergenic" for regions between the two EST matches. A

Table 1. List of all the Gene Finders Applied to the Araset Sequences. All Systems have Been Trained for *A. thaliana*

Program	Reference	Version
Genscan	1	Data from [14]
GlimmerA	33	1.0
GeneMarkHMM	8	2.2a
FGenesP	Solovyev (1997), unpublished	Data from [14]
FGenesH	Salamov, Solovyev (1991), unpublished	1.0.
FGenesHGC	Unpublished, can deal with GC splice sites	
AUGUSTUS	18	From web site
EuGène	7	3.5

Table 2. Accuracy of the Different Gene Finders on Araset Given by Exon Level Sensitivity and Specificity (Sne/Spe) and Gene Level Sensitivity and Specificity (Sng/Spg). Two specificities are given to allow comparison with AUGUSTUS in which Spg is computed with the evaluation procedure provided with Araset, but ignores all predictions in the first and last 2 kb of every sequence (context). Spg' is computed on the complete sequence (the only measure available on AUGUSTUS web site). The simple *ab initio* variant of EuGène is in agreement with AUGUSTUS, the best *ab initio* gene finder. The introduction of additional evidence clearly increases prediction quality with the most informed variant correctly predicting more than three genes out of four.

Program	Sne	Spe	Sng	Spg	Spg'
Genscan	63	70	17	19	-
FGenesP	42	59	6	11	-
GeneMarkHMM	83	78	41	37	-
GlimmerA	67	67	30	19	-
FGenesH	88	84	56	53	-
FGenesHGC	88	88	57	55	-
AUGUSTUS (*ab initio*)	89	-	62	-	39
EuGène (*ab initio*)	83	87	62	59	38
EuGène (EST)	87	88	71	66	41
EuGène (Protein)	90	89	74	69	44
EuGène (Full)	91	91	77	73	45

similarly extreme vote is issued for an intergenic region immediately before and after the region between the two EST matches, forcing the gene to lie between these boundaries. The flexibility of EuGène was crucial here. The corresponding annotations have been used by The Institute for Genome Research and compared to The Arabidopsis Information Resource (TAIR version 5). Predictions that did not appear in TAIR were experimentally tested by rapid amplification of cDNA ends and hundreds of new genes predicted by EuGène and TwinScan were added to TAIR (version 6; [19]). Hundreds of more genes identified by EuGène have been tested and integrated since [20].

4.2. Eugène Used to Annotate *Medicago truncatula*

The genome of *Medicago truncatula* (barrel medic), a model organism for legumes, is currently being sequenced and annotated by the international consortium IMGAG (http://www.medicago.org/genome/IMGAG). Although the genomes of *A. thaliana* and *M. truncatula* are not very distant, the training of genome-specific versions of gene finders greatly improves gene predictions. Furthermore, the development of new trainable splice site prediction tools, such as SpliceMachine [10], offered an opportunity to incorporate better site predictions into EuGène. Despite different technologies and score semantics, the incorporation of SpliceMachine scores instead of NetStart/NetGene2 and SplicePredictor required minor work. A first *ab initio* variant of EuGène was designed based on SpliceMachine (splice and ATG sites) and GHMMs. These mathematical models were estimated on a first reliable set of genes built from EST clusters [41]. This basic *ab initio* variant was enhanced by the integration of an N-terminal targeting sequence detection software tool, Predotar [42]. Indeed, targeting sequences are part of the coding sequences, but are often classified as noncoding by GHMMs because of their strong compositional biases. The peptide header detection allowed voting for nearby ATGs. The corresponding rescaling parameters with refined functions have been estimated with a second set of independently curated genes (all available at http://medicago.toulouse.inra.fr/Mt/GLIP/).

Simultaneously, a specific version of FGenesH was built for *M. truncatula* by the SoftBerry company (http://www.softberry.com/). This gene finder predicts complete genes that can be decomposed into regions (exon/intron) and signals (ATG/STOP and splice sites). A general plug-in capable of integrating such information was designed. Because the prediction lacks associated scores, a constant vote for each of the corresponding regions and signal types is used. These constants are estimated as other rescaling parameters by maximizing the prediction quality on the same expert data set. Integration of FGenesH predictions yields another variant of EuGène that can still be considered as *ab initio*. It is interesting to note that this version is much better than either of its components (see Table **2**). Besides ESTs of *M. truncatula*, those of other legumes, such as *Lotus japonicus*, *Pisum sativum* (pea), and *Glycine max* (soybean) were also used as extrinsic sources of information.

Considering protein similarities, a peptide databank derived from *M. truncatula* EST clusters was used to improve the detection of new members of gene families. In other cases, other reliable protein databanks were used such as SwissProt, *A. thaliana* proteome (TAIR 6), and ProDom (all protein domains with more than two elements). The latter databank is derived from SP/TrEMBL, but a lot of hits with spurious sequences are avoided thanks to the nonsingleton domains.

Conserved regions between two genomic sequences representing likely coding exons can be detected with TblastX. The integration of such information into EuGène has previously been described [21]. The associated plug-in votes for coding regions when protein-level similarities are observed. For its completeness and maturity, the whole *Arabidopsis* genome was also utilized.

In contrast to previous genome annotations that often rely on different pipelines for each pseudo-molecule or sequencing center, EuGène allows the production of a single, uniform, and automatic consensus prediction that includes information from a variety of sources. Before applying this recipe for the annotation of *M. truncatula*, different gene finding systems were evaluated on a set of 172 new well-curated genes built by mapping transcripts to freshly sequenced bacterial artificial chromosomes (Table **3**).

4.3. EuGène Used to Annotate *Botrytis cinerea*

Botrytis cinerea is a filamentous fungus responsible for the grey mould disease, affecting more than 200 host plants and inflicting serious crop losses worldwide. It also provokes the famous noble rot that gives its characteristic flavor to Sauternes wine (sweet Bordeaux). *B. cinerea* genome has been sequenced recently by Genoscope (http://www.cns.fr) at 10.5x coverage. Compared to the genome of plants, that of *Botrytis* is much smaller (40 Mb) with small introns [43]. With EST clusters with significant similarity to fungal proteins, a set of 446 curated genes was built from which 300 were extracted: 100 were used for training statistical models for EuGène (ATG and splice site prediction with SpliceMachine and GHMM estimation) and submitted to SoftBerry for training FGenesH (www.soft-berry.com), 100 different ones for parameter rescaling in EuGène, and the 100 remaining for evaluation.

Table 3. Results of Different Gene Finders Applied to Fresh Gene Sequences with curated Annotation. Except for EuGène and FGenesH, trained versions of *Arabidopsis* were used, because no version was available for *M. truncatula*. Interestingly, making FGenesH predictions available to EuGène allows to outperform both gene finders. Integration of additional EST and protein similarities allows correct prediction of four out of five genes, a very optimistic evaluation of the actual performance on genomic sequences because the curated dataset is associated with lot of evidence. The genes wrongly predicted (20%) received mostly an incorrect ATG codon, as expected because the lack of experimental evidence obstructs the choice of a "correct" ATG. Training and test sequences are available at http://medicago.toulouse.inra.fr/Mt/-GLIP/

Gene finder	Sne	Spe	Sng	Spg
Genscan (*A. thaliana*)	69.6	78	25.8	29
GeneMark.HMM (*A. thaliana*)	73.1	76.6	32.4	31.6
FGenesH (*A. thaliana*)	85.3	81.4	47	46.5
FGenesH (*M. truncatula*)	85.1	80.7	52.8	47.8
EuGène (*ab initio, M truncatula*)	84.7	85.4	55.5	50.5
+ FGenesH	90	86.9	63.2	56.4
+ Protein similarity	92.4	88	69.2	61.8
+ Transcript similarity	94.4	94.6	80.2	79.4

Table 4. EuGène Variants Compared to FGenesH on 100 Curated Gene Sequences. All gene finders were trained with the same gene set of *B. cinerea*. on this rather compact genome, with short introns, the *ab Initio* variant of EuGène already offers an excellent prediction accuracy. The FGenesH prediction available to EuGène increases in the accuracy. The integration of all similarity based evidence (cognate protein have been removed for the analysis) allows correct prediction of more than nine out of ten genes.

Gene finder	Sne	Spe	Sng	Spg
FGenesH	86.2	80.7	73.6	60.2
EuGène (ab initio)	89.3	92.2	75.3	74.5
+ FGenesH	94.7	96.3	87.1	87.1
+ Protein similarity	96.9	97.2	90.1	90.1
+ EST similarity	97.8	97.5	92.1	92.1

As before, of several variants of EuGène tested, the first is a purely *ab initio* one that uses only SpliceMachine and GHMMs. As for *M. truncatula*, the output from FGenesH was incorporated to define a second variant. Then, we searched for protein similarity using three databanks, namely one of fungal proteins, SwissProt, and TrEMBL. Fungal proteins previously considered to build the curated gene set were removed from the similarity analysis to avoid a bias in the process. Finally, similarities to ESTs from *B. cinerea* and the closely related *Sclerotinia sclerotiorum* genome (www.broad.mit.edu/annotation/genome/sclerotinia_sclerotium) were added. Incorporation of additional sources of evidence, including existing gene predictions, is effectively useful for the quality of gene prediction obtained with EuGène (Table **4**).

In addition to these three organisms, EuGène has been or is currently being trained and applied to large-scale annotation of a number of organisms, including *Populus trichocarpa* (poplar) [44], *Solanum lycopersicum* (tomato), *Physcomitrella* (moss), *Ostreococcus tauri* (a green alga) [45], *O. lucimarinus* [46], *Micromonas pusilla* (a green alga), *Ectocarpus siliculosus* (a brown alga), *Laccaria bicolor* (a fungus), *Melampsora larici-populina* (a fungus), *Meloidogyne incognita* (a nematode), *Oryza sativa* (rice), and will be trained on *Arabidopsis lyrata*, *Capsella rubella*, and *Eucalyptus globulus* as well as other genome projects currently in the launching phase.

5. CONCLUSIONS

For several years, protein gene prediction approaches have been classified as either intrinsic (or *ab initio*, based on statistical properties), or extrinsic (based on similarities and conservation). After some hybrid gene finding systems, such as GenomScan [13], TwinScan [3], or even the recent NSCAN [4], the general trend is toward highly integrative gene finders. In the plant kingdom, this track has been followed by EuGène for several years and has been applied to several genome-scale sequencing and annotation projects.

Integrative gene finding is a complex problem of information fusion that will continue to evolve in the future. Several issues remain open, such as easier handling of vertical and horizontal correlations between sources of evidence, choice of the criteria for parameter estimation, better incorporation of existing sources of evidence, integration of new sources of evidence, and prediction of new region types (e.g. pseudogenes, noncoding RNA genes, and promoters). In the near future, genome sequencing will become more affordable and will allow a wider exploration of the tree of life. A major issue will be adaptation of gene prediction to genomes of very opposite style, from very compact to very large, with strongly documented to barely known resources.

Also very important will be to facilitate the actual training itself. Most of the actual systems require some manual tweaking based on expertise and construction of curated data sets as well that can be both time consuming and strongly influence the quality of the final system. Even more complex are integrative systems that should select the additional sources of information and incorporate them optimally for the organism considered. A complete gene finding system should automate all these aspects as it has already been done to some extent for pure *ab initio* gene finders, such as Genemark.ES [47].

ACKNOWLEDGEMENTS

We would like to thank C. Town and F. Cheung from The Institute for Genome Research for providing the FGenesH predictions and PASA-based EST alignments [48] for *Medicago truncatula* annotation in the framework of the IMGAG consortium. This work was supported by a grant from the European Union Sixth Framework Programme on Food Safety and Quality (FOOD-CT2006-016214).

REFERENCES

[1] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **1997**; 268: 78-94.

[2] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **1989**; 77: 257-286.

[3] Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics* **2001**; 17 Suppl 1: p. S140-8.

[4] Brown RH, Gross SS, Brent MR. Begin at the beginning: predicting genes with 5' UTRs. *Genome Res* **2005**; 15: 742-747.

[5] Decaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE. Conrad: gene prediction using conditional random fields. *Genome Res* **2007**; 17(9): 1389-1398.

[6] Allen JE, Salzberg SL. Jigsaw: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **2005**; 21: 3596-3603.

[7] Schiex T, Moisan A, Rouzé P. EuGène: an eucaryotic gene finder that combines several sources of evidence. *Lect Notes Comput Sci* **2001**; 2066: 111-125.

[8] Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A. Detection of new genes in a bacterial genome using markov models for three gene classes. *Nucleic Acids Res* **1995**; 23: 3554-3562.

[9] Zhang MQ, Marr TG. A weight array method for splicing signal analysis. *Comput Appl Biosci* **1993**; 9: 499-509.

[10] Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y. Splicemachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **2005**, 21: 1332-1338.

[11] Schlueter SD, Dong Q, Brendel V. Geneseqer@plantgdb: gene structure prediction in plant genomes. *Nucleic Acids Res* **2003**; 31: 3597-3600.

[12] Gremme G, Brendel V, Sparks M, Kurtz S. Engineering a software tool for gene structure prediction in higher organisms. *Inf Softw Technol* **2005**; 47: 965-978.

[13] Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Res* **2001**; 11: 803-816.

[14] Gross SS, Brent MR. Using multiple alignments to improve gene prediction. *J Comput Biol* **2006** 13: 379-393.

[15] Alexandersson M, Cawley S, Pachter L. Slam: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res* **2003**; 13: 496-502.

[16] Sarawagi S, Cohen W. Semi-Markov conditional random fields for information extraction. In: Saul LK, Weiss Y, Bottou Y Eds, Advances in Neural Information Processing Systems, Proceedings of the 17th Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, **2005**; 17: 1185-1192 [http//books.nips.cc/nips17.-html]

[17] Bellman R, Kalaba R. Dynamic programming and modern control theory. Acadamic Press London 1965.

[18] Stanke M, Steinkamp R, Waack S, Morgenstern B. Augustus: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **2004**; 32: W309-12.

[19] Moskal WAJ, Wu HC, Underwood BA, Wang W, Town CD, Xiao Y. Experimental validation of novel genes predicted in the un-annotated regions of the arabidopsis genome. *BMC Genomics* **2007**; 8: 18.

[20] Aubourg S, Martin-Vigniette ML, Brunaud V, *et al*. Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics* **2007**; 8: 401.

[21] Foissac S, Bardou P, Moisan A, Cros M, Schiex T. EuGene'Hom: a generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res* **2003**; 31: 3742-3745.

[22] Foissac S, Schiex T. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* **2005**; 6: 25.

[23] Birney E. Ensembl: a genome infrastructure. *Cold Spring Harb Symp Quant Biol* **2003**; 68: 213-215.

[24] Howe KL, Chothia T, Durbin R. Gaze: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res* **2002**; 12: 1418-1427.

[25] Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **2006**; 7: 62.

[26] Majoros WH, Salzberg SL. An empirical analysis of training protocols for probabilistic gene finders. *BMC Bioinformatics* **2004**; 5: 206.

[27] Allen JE, Majoros WH, Pertea M, Salzberg SL. Jigsaw, Genezilla, and GlimmerHMM: puzzling out the features of human genes in the encode regions. *Genome Biol* **2006**; 7 Suppl 1: S9.1-13.

[28] Stanke M, Tzvetkova A, Morgenstern B. Augustus at EGasp: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **2006**; 7 Suppl 1: S11.1-8.

[29] Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics* **2006**; 7: 447.

[30] Pedersen AG, Nielsen H. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proc Int Conf Intell Syst Mol Biol* **1997**; 5: 226-233.

[31] Hebsgaard SM, Korning PG, Tolstrup N, *et al*. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* **1996**; 24: 3439-3452.

[32] Kleffe J, Hermann K, Vahrson W, Wittig B, Brendel V. Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucleic Acids Res* **1996**; 24: 4709-4718.

[33] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with Glimmer. *Nucleic Acids Res* **1999**; 27: 4636-4641.

[34] Pavy N, Rombauts S, Déhais P, *et al*. Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* **1999**; 15: 887-899.

[35] Korning PG, Hebsgaard SM, Rouzé P, Brunak S. Cleaning the Genbank *Arabidopsis thaliana* data set. *Nucleic Acids Res* **1996**; 24: 316-320.

[36] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **1998**; 8: 967-974.

[37] Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-3402.

[38] Crowe ML, Serizet C, Thareau V, *et al*. CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res* **2003**; 31: 156-158.

[39] Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2004 http: //www.repeatmasker.org (accessed on Jan. 2008).

[40] Seki M, Narusaka M, Kamiya A, *et al*. Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **2002**; 296: 141-145.

[41] Journet E, van Tuinen D, Gouzy J, *et al*. Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. *Nucleic Acids Res* **2002**; 30: 5579-5592.

[42] Small I, Peeters N, Legeai F, Lurin C. Predotar: a tool for rapidly screening proteomes for n-terminal targeting sequences. *Proteomics* **2004**; 4: 1581-1590.

[43] Fillinger S, Amselem J, Artiguenave F, *et al*. The genome projects of the plant pathogenic fungi *Botrytis cinerea* and *Sclerotinia sclerotiorum*. In: Jeandet P, Clément C, Conreux A Eds, Macromolecules and Secondary Metabolties of Grapevine and Wine. Technique & Documentation, Paris, 2007; 125-133.

[44] Tuskan GA, Difazio S, Jansson S, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **2006**; 313: 1596-1604.

[45] Derelle E, Ferraz C, Rombauts S, *et al*. Genome analysis of the smallest free-living eukaryote ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci USA* **2006**; 103: 11647-11652.

[46] Palenik B, Grimwood J, Aerts A, *et al*. The tiny eukaryote ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci USA* **2007**; 104: 7705-7710.

[47] Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **2005**; 33: 6494-6506.

[48] Haas BJ, Delcher AL, Mount SM, *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **2003**; 31: 5654-5666.

Received: November 09, 2007 Revised: February 12, 2008 Accepted: March 03, 2008