BioTechnologia vol. 93(3) • pp. 318-332 • 2012
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPER

Annotating a non-model plant genome – a study on the narrow-leafed lupin

ANDRZEJ ZIELEZIŃSKI [1], PIOTR POTARZYCKI [1], MICHAŁ KSIĄŻKIEWICZ [2], WOJCIECH M. KARŁOWSKI [1]*

[1] Laboratory of Computational Genomics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland

[2] Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland

* Corresponding author: wmk@amu.edu.pl

Abstract

We present here a highly portable and easy-to-use gene annotation system CEL (Computational Environment for annotation of Legume genomes) that can be used to annotate any type of genomic sequence – from BAC ends to complete chromosomes. CEL's core engine is modular and hierarchically organized with an open-source structure, permitting maximum customization – users can assemble an individualized annotation pipeline by selecting computational components that best suit their annotation needs. The tool is designed to speed up genomic analyses and features an algorithm that substitutes for a biologist's expertise at various steps of gene structure prediction. This allows more complete automation of the labor-intensive and time-consuming annotation process. The system collects and prioritizes multiple sources of *de novo* gene predictions and gene expression evidence according to the confidence value of underlying supporting evidence, as a result producing high-quality gene-model sets. The data produced by CEL pipeline is suitable for direct visualization in any genome browser tool that supports GFF annotation format (e.g. Apollo, Artemis, Genome Browser etc.). This provides an easy means to view and edit individual contigs and BACs using just mouse's clicks and drag-and-drop features. Finally, we show that CEL produces accurate annotations for novel draft genomes,of low quality or mostly non-existent, as in the case of narrow-leafed lupin where the training-data are limited.

Key words: genome annotation, pipeline, software, narrow-leafed lupin, draft-genome, plant genome

Introduction

Genomic sequences are rapidly being published for thousands of species and they represent a milestone in research on the biology of any particular organism. However, the complete DNA sequence alone represents only the first step in elucidation of the encoded biological function. Therefore, the value of the sequencing efforts is directly linked with the accuracy of DNA annotation – a detailed description of all genetic elements, in particular, genes and their products. Such information along with subsequent experimental work allow deciphering of mechanisms involved in functioning of the organism as well as interpretation necessary to extract their biological significance in the context of complete biological processes (Stein, 2001).

The recent availability of high-throughput DNA sequencing technologies have led to a situation where new genomes are being sequenced faster than they could be annotated. As of 2012, there are already 221 completely sequenced, but unpublished genomes, and more than 750 Eukaryotic genome projects under way (Pagani et al., 2012). Genome annotation itself is a complex, multi-step process which requires a human-curated integration of diverse sources of computational evidence, including results from *ab initio* prediction programs as well as homology-based searches.

In practice, a limitation of the manually-curated multiple-evidence approach is the need to combine computational results from a disparate set of independent annotation programs. No equivalence in outputs of such software makes cooperative data analysis very difficult for a non-bioinformatics user.

In addition, these tools are often designed to work on a single contiguous sequence (contig) at a time, while many annotation efforts require the analysis of thousands of assembled contigs.

The automatic annotation pipelines are created in most of the major genome sequencing projects. Un-

fortunately, they are often focused on selected, single or a group of related organisms by including specific prediction methods and providing hard-encoded reference data. Other approaches, involving creation of more general computational tools, lead to the development of complicated and very often user-unfriendly systems (Liang et al., 2009; International Rice Genome Sequencing Project, 2005).

Thus, despite the best efforts of the specialized bioinformatics communities, large numbers of unannotated genomes continue to accumulate, underscoring an urgent need for simpler, more portable annotation strategies.

Here, we start by reviewing the existing approaches commonly used to predict genes in eukaryotic genomes and underline their intrinsic advantages and limitations. As a proof-of-concept, we also present a new, easy-to-use automatic, integrated, comprehensive computational environment (CEL) dedicated to the annotation and comparative analysis of genomic sequences, particularly useful for non-model plant genomes. Our annotation pipeline integrates the results from multiple programs and facilitates an optional human curation of computational data. The simple and modular design allows tailoring of each program for an individual step in the annotation process, and it can be used independently of all other programs in the package. Such a design strategy allows users to assemble an individualized annotation pipeline by selecting those computational components that are most appropriate to their annotation needs. The connection between components of the pipeline is achieved by the translation of computational evidence from the native annotation program output into the standardized format (General Feature Format; GFF). The GFF file format facilitates an integration of multiple computational results. It can be directly curated and modified by any biologist using standard sequence editing and visualization tools such as Apollo (Lewis et al., 2001), Artemis (Carver et al., 2012), GBrowse (Donlin et al., 2007), the UCSC genome browser (Kent et al., 2002) or the Ensembl Genome Browser (Kersey et al., 2010). Moreover, the design of our pipeline allows the user to quickly look at the results at any time during the computations, before all completion of the annotation process.

To demonstrate the potential of our new tool, we present results of the annotation of available genomic sequences for narrow-leafed lupin (NLL). Lupins, members of the legume family, are a valuable crop because the grains are high in protein and fiber and low in starch and oil (Erba et al., 2005) and are becoming recognized as a potential human health food (Lee et al., 2005; Duranti et al., 2008). However, agronomically very important, lupins still lack extensive genomics resources and studies. Currently, plants are being subjected to intensive genetic analysis including linkage mapping and genomic library development (Nelson et al., 2010; Gao et al., 2011). The NLL genome is of medium size (900 Mbp) and composed of 40 chromosomes. The ongoing genome sequencing project and efforts to characterize selected parts of NLL genome provide an opportunity to take "a first look" into its composition and organization.

Most of the effective annotation tools depend on the comparison with already available molecular data. It is therefore crucial for the quality of annotation results to include all of the available data that can be used for the computations. Using NLL genomic DNA as a case study, we additionally present here a list of available sequence resources useful for its annotation. Such information can be compiled for any species and be used in our annotation pipeline to extract functional information from non-model, not-yet-analyzed genomes.

Overview of major computational challenges of genome annotation

The collective process of identifying genes (structural annotation) and assigning a function to each of them (functional annotation) is commonly referred to as genome annotation. Protein-coding genes, which for the most part dictate the biological function, comprise a small fraction of higher eukaryotic genomes, <30% of Arabidopsis genome (The Arabidopsis Information Resource, TAIR) and even a smaller part of the human genome (<3%) (Lander et al., 2001). This makes the identification of coding sequences (CDSs) in the ocean of non-coding sequences extremely difficult. Additionally, in eukaryotes, coding regions (exons) are often widely interspersed with non-coding intervening sequences (introns). For instances, the human dystrophin gene is composed in 99% of introns, some of which are 100 kb in size (Sleator, 2010). On the other hand, some Arabidopsis genes contain exons which are only 3-bp long (Mathé et al., 2002). Short sequences of this type are usually beyond the detection limit of available computa-

Table 1. Software commonly used for genome annotation

Ab initio and evidence-drivable gene predictors		
Augustus	http://bioinf.uni-greifswald.de/augustus/	uses EST-based and protein-based evidence hints in the *de novo* gen prediction. Highly accurate
GENSCAN	http://genes.mit.edu/GENSCAN.html	hidden Markov models HMMs-based gene predictor. The online service trained for predicting genes of vertebrates, Arabidopsis thaliana and maize
Geneid	http://genome.crg.es/software/geneid/	highly configurable gene predictor. Supports integrating predictions (via GFF) from multiple sources (ESTs, BLAST best-scored hits)
GeneMark	http://exon.gatech.edu/	self-training program (just 10MB sequence is needed for training), supports numerous eukaryotic and prokaryotic genomes
EST, protein and RNA-seq aligners and assemblers		
BLAST	http://blast.ncbi.nlm.nih.gov/	compares a query sequence with those contained in nucleotide and protein databases using Karlin–Altschul statistics
BLAT	http://genome.ucsc.edu/	500 times faster than popular existing tools for mRNA/DNA alignments and 50 times faster for protein alignments. May miss more divergent sequences
Sim4	http://pbil.univ-lyon1.fr/	splice-aware cDNA-to-DNA alignment tool
SplicePredictor	http://bioservices.usd.edu/splicepredictor/	splice-site-aware alignment algorithm that can align both protein and EST sequences to a genome
Splign	http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi	computes cDNA-to-DNA splicealignments with identification of paralogs
Cap3	http://pbil.univ-lyon1.fr/cap3.php/	DNA sequence assembly program generates a consensus sequences. Uses forward-reverse constraints to correct assembly errors and link contigs
crossmatch	http://www.phrap.org/	compares a set of reads to a set of vector sequences and produces vector-masked versions of the reads
Protein level annotation		
InterProScan	http://www.ebi.ac.uk/InterProScan/	search for domains/motifs in the InterPro database
Pfam	http://pfam.sanger.ac.uk/	analyze a protein sequence for Pfam domain/family matches
HMMER 3.0	http://hmmer.janelia.org/	uses HMMs to search sequence databases for homologs of protein sequences. Highly accurate and able to detect remote homologs
GOAnno	http://bips.u-strasbg.fr/GOAnno/	BLAST search on the Gene Ontology database
COGNITOR	http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html	compares a query sequence to the COG (Cluster of Orthologous Groups of proteins) database.
ReviGO	http://revigo.irb.hr/	summarizes long, unintelligible lists of GO terms by finding a representative subset of the terms using a simple clustering algorithm
Genome browsers for curation		
Apollo	http://apollo.berkeleybop.org/	java-based genome browser that allows the user to create and edit gene models and write their edits to a remote database
Artemis	http://www.sanger.ac.uk/resources/software/artemis	java-based integrated platform for visualization and analysis of sequence features and high-throughput sequence-based experimental data
JBROWSE	http://jbrowse.org/	fast, modern genome browser written primarily in JavaScript with a fully dynamic AJAX interface. Perfect for Web-based use

tional methods and many programs simply ignore such short exons. The extreme variation in both, the number and length of intron and exon sequences is particularly difficult during the detection of exon borders. What complicates the identification of protein-coding genes and their correct genomic structure further, is the high frequency of alternative splicing in most eukaryotic genes. It is estimated that more than 50 and 95% of intron-containing genes in Arabidopsis (Marquez et al., 2012) and humans (Sleator, 2010) respectively, show evidence of at least one alternative gene variant. Other factors complicating genome prediction include the presence of overlapping genes, though rare in eukaryotic genomes, there are some documented cases in plants (Quesada et al., 1999) and animals (Makalowska et al., 2005). In addition to protein-coding genes, a large proportion of the eukaryotic genomes encodes functional RNA sequences that play an important role in the regulation of eukaryote gene expression (Ebert et al., 2012). In the literature there are regular reports on some new, unexpected and non-canonical cases of functional genetic elements which additionally increase the level of difficulty of the genome annotation problem, and currently there is no such program that can cover all the possible variants. Thus, defining the precise start and stop position of a gene and the splicing pattern of its exons among all the genuine non-coding sequence is still one of the major bioinformatics challenges.

The protein-coding genes in genomic DNA sequences are identified using two general approaches: *ab initio* gene prediction methods (intrinsic methods) and similarity methods (extrinsic methods).

Ab initio gene predictors owe their name to the fact that they use mathematical models rather than external evidence (such as DNA and protein sequence alignments) to identify genes and to determine their intron-exon structures. They rely on the intrinsic features of the DNA sequence to discriminate between the coding and non-coding regions, allowing the identification of genes by detecting signals known as typical of gene structures (promoters, termination signals, splicing signals and junction boundaries). These methods need to be trained on a set of known genes assuming that the genes within a genome share similar compositional properties that are species specific. There is a number of well-tested groups of such programs widely used for gene prediction (Tab. 1). With enough training data, the gene-level sensitivity of *ab initio* tools can approach 100%. However, the accuracy of the predicted intron-exon structures is usually much lower, ~60-70% (Yandell and Ence, 2012) and drops drastically (~30%) if the algorithm is required to predict the entire gene structure correctly (Stein, 2001). This means that most of the genes predicted by *ab initio* programs contain errors ranging from an incorrect exon boundary to a missed or phantom exon. A major limitation in usability of this type of gene identification programs is the need to provide a set of known, well-characterized genes that can be used for training the algorithm for each organism. However, in most cases, no training is available for genomes of non-model species. And in such cases, gene prediction models fine-tuned to the closest phylogenetic group are most often used. The gene prediction process might be further complicated by the software implementation details, which may lead to a situation where different *ab initio* prediction programs that are based on the same algorithm and trained for the same organism produce different predictions. However, the most important factor for the algorithm performance is the content of gene samples in the training set, thus the final accuracy of gene prediction depends on the quality of source dataset. In this light, in order to optimize gene prediction methods, a common practice is to combine the predictions of several programs in order to obtain a consensus result (Yok and Rosen, 2011) where only exons detected by two or more predictions are kept for further annotation steps.

The *ab initio* gene finding programs derive full gene models from DNA data based solely on knowledge of the sequence features associated with the protein coding domain. The similarity of a region of the genome to a sequence that is already known to be transcribed and translated cannot only refine the exon-intron boundaries of gene models but also provide evidence that computationally predicted genes are actually expressed. The basic tools for detecting similarity between sequences are local alignment methods ranging from the optimal Smith-Waterman algorithm (Smith and Waterman, 1981) to fast heuristic approaches implemented in programs like FASTA and BLAST (Pearson and Lipman, 1988; Altschul et al., 1997). A statistically significant match (measured by *p*-value or *e*-value) to a cDNA/EST sequence or even a *in-silico*-translated match to a gene from other species might be good evidence that an investigated region be-

longs to an expressed gene. The EST/cDNAs sequences are the most relevant information used to establish the structure of a gene, especially if they come from the same organism as the genome to be annotated. In general, the programs that use gene expression data have the advantage of generating fewer false predictions than *ab initio* methods. However, the lack of predictions with the similarity-based-method does not imply that the gene is absent. It might result from incomplete data set used for the analysis. This is very often a case of genes that are expressed either under very specific conditions or at a low level. Another scenario represents novel genes that show limited similarity to sequences available in databases. It has been estimated that only a half or less of genes can be annotated by searching for similarities to other known genes or proteins and the remaining genes need to be identified using *de novo* approaches (Mathé et al., 2002).

The commonly used approach during the annotation process is to combine evidence from both *ab initio* gene predicting programs as well as sequence similarity searches against databases of previously identified proteins and expressed RNA. This technique is also used in the annotation pipeline presented here. However, certain genetic features, like transposable elements (TE), may additionally complicate the annotation procedure. Millions of copies of TEs cover a large proportion of eukaryotic genomes: 50% of the 3.2 Gb human genome (de Koning et al., 2011), and more than 80% of the 17 Gb bread wheat genome (Cantu et al., 2010). *Ab initio* predictors examine sequences which search for nucleotide motifs that occur more commonly than expected by chance, consequently, often annotate these TEs as genes. In addition, most TE genes are expressed and represented in cDNA libraries, therefore, searches for sequence similarities will also indicate that TEs are transcribed and may be considered as a gene. Since these false positive gene predictions cannot be distinguished by conventional gene prediction methods alone, annotators, in order to remove them from the gene candidate list, look for such elements by comparing query sequences with those in curated sequence repeats libraries.

In conclusion, the gene annotation is a lengthy, time-consuming and recursive process. Only ten versions of the Arabidopsis genome have been released so far (The Arabidopsis Information Resource, TAIR). Obviously, the difficulty of the process increases along with the size and complexity of the genome organization. It requires careful chaining of numerous programs, algorithms and methods under the supervision of an expert biologist. However, most of the programs produce results in a specific, not-compatible format, which makes manual data analysis extremely difficult. Hence, there is an increased demand for a user-friendly, easy-configurable software that minimizes the unnecessary manual curation by automating this extensive and laborious analysis.

Results

Implementation of the annotation system

The CEL pipeline automatically combines the output from alignment-based evidences with *ab initio* gene prediction results using user-defined parameters to obtain a final set of gene annotations.

A number of useful annotation pipelines have been developed by genome annotation communities to analyze plant (International Rice Genome Sequencing Project, 2005; Liang et al., 2009) and animal genomes (Flicek et al., 2012). However, in general, such procedures are most often based on massive informatics and solution-specific resources, which makes them inaccessible for outside users. When available, the setup of such systems is very complex and the installation of various tools may require extensive skills in computer science. In contrast, our aim was to develop a pre-compiled, ready-to-use package of a CEL annotation system that contains all required bioinformatics tools (publicly available and free for academic use) and is easy to install on most common platforms (UNIX, Linux, OS X, Windows).

Input source data

The CEL annotation system requires the source genomic sequence in FASTA format (Lipman and Pearson, 1985) and a configuration file (for details see README files within the software package) describing sequence database locations, and various compute parameters.

The pipeline is designed to allow processing of various genomic sequences of any length – from short BAC ends (BES) to large, completed BAC clones and draft genomes. The input sequences are automatically fragmented into series of chunks (of default size of 1 kbp). Each fragment is then separately computationally analyzed and the results of calculations for all the chunks are merged. In this way, genomic DNA fragments of any size can be annotated even on a laptop computer.

The reference sequence datasets (cDNA libraries, proteins and genes) can be provided as a list of file names and/or Internet sources (HTTP or FTP) from which sequence records will be automatically downloaded. This allows analysis of incomplete partial genome assemblies and independent annotation of regions of interest by using custom selected data sets.

Whichever annotation strategy (see overview of annotation methods and pitfalls described above) is selected, the accuracy of annotation inevitably depends on the amount and quality of reference datasets. Thus, before sequences are submitted to the annotation workflow, they are first subjected to a quality check (i.e. duplicate and empty sequence records are filtered out and ambiguous non-IUPAC characters are masked) to create a high quality non-redundant sequence data sets collectively for a specific type of sequence (cDNA libraries, full-length mRNAs, proteins, etc.) for a given species.

Modular architecture of CEL

The architecture of the pipeline is modular and fully customizable using a single configuration file. This makes it fairly easy to modify the general annotation workflow as well as specific parameters for each of the individual analysis programs. It is is especially useful when working with a genome that has not been yet analyzed, and thus appropriate annotation parameters and reference sequence resources have not been set. The analysis pipeline is divided into five steps (Fig. 1): I – repeat sequences annotation and masking, II – coding sequence detection and determination of exon-intron boundaries, III – gene models determination, IV – functional annotation and V – visualization and manila expert curation.

I. Repeat sequence annotation and masking

Unless repeat sequences are effectively excluded from the source sequence, the resulting gene models will most probably contain portions of TEs and viruses. Two strategies are implemented in the CEL pipeline to identify the repeats. First, RepeatMasker (Saha et al., 2008) is used to scan the input sequence for low-complexity regions (process called soft-masking). Lower-case masking of such sequences is a signal for BLAST and other alignment-based tools that these regions should be treated as repeats. The second step of repetitive sequence identification includes similarity search at the protein level against RepBase dataset (Jurka et al., 2005). The RepBase is a well-curated library of known repeat families from diverse eukaryotic organisms. It has been shown that such a two-way approach greatly enhances repeat identification in both well-characterized and unannotated genomes (Smith et al., 2007).

II. Detection of coding sequences (CDS) and determination of exon-intron boundaries

Gene structures are predicted by a CEL pipeline using *ab initio*, sequence similarity and combination of thereof approaches.

Two *ab inito* predictors, GENSCAN (Burge and Karlin, 2007) and Augustus (Stanke and Morgenstern, 2005) are integrated by default to predict intron-exon gene models. In overall performance GENSCAN detects the coding regions with high sensitivity and specificity values that reach 91 and 92% for Arabidopsis, and 96 and 93% in the case of fruit fly. Augustus includes additional mechanisms to incorporate extrinsic data into the *ab initio* gene prediction framework to improve its accuracy. CEL's modular architecture means that any gene predictor can be integrated into its structure with minimal modification of the software code.

BLAST is used for identification of known cDNAs and proteins with significant similarity to the input genomic sequence. BLAST is not well suited for prediction of splice site boundaries and in most cases its alignments are only rough approximations of CDS prediction (Slater and Birney, 2005). CEL annotate pipeline realigns the sequences identified by BLAST using additional alignment algorithms to achieve higher precision of exon boundaries prediction (described later in more detail).

The most useful, yet of lower quality, in genome annotation are EST sequences. Therefore, mapping ESTs to genomic sequence is preceded by three steps. The EST pre-processing includes identification and elimination of vector sequences using Univec database and a cross-match program (Green, unpublished data, http://phrap.org/). The sequences are clustered into larger contigs using CAP3 (Huang and Madan, 1999). Finally, the gene structure prediction step includes the use of two splice-aware alignment tools, sim4 (Florea et al., 1998) and SplicePredictor (Brendel et al., 2004). Sim4 is one of the most frequently used programs for studying gene-to-genome alignment and alternative splicing. The SpilicePredictor algorithm is designed to tolerate a high percentage of mismatches and insertions or deletions

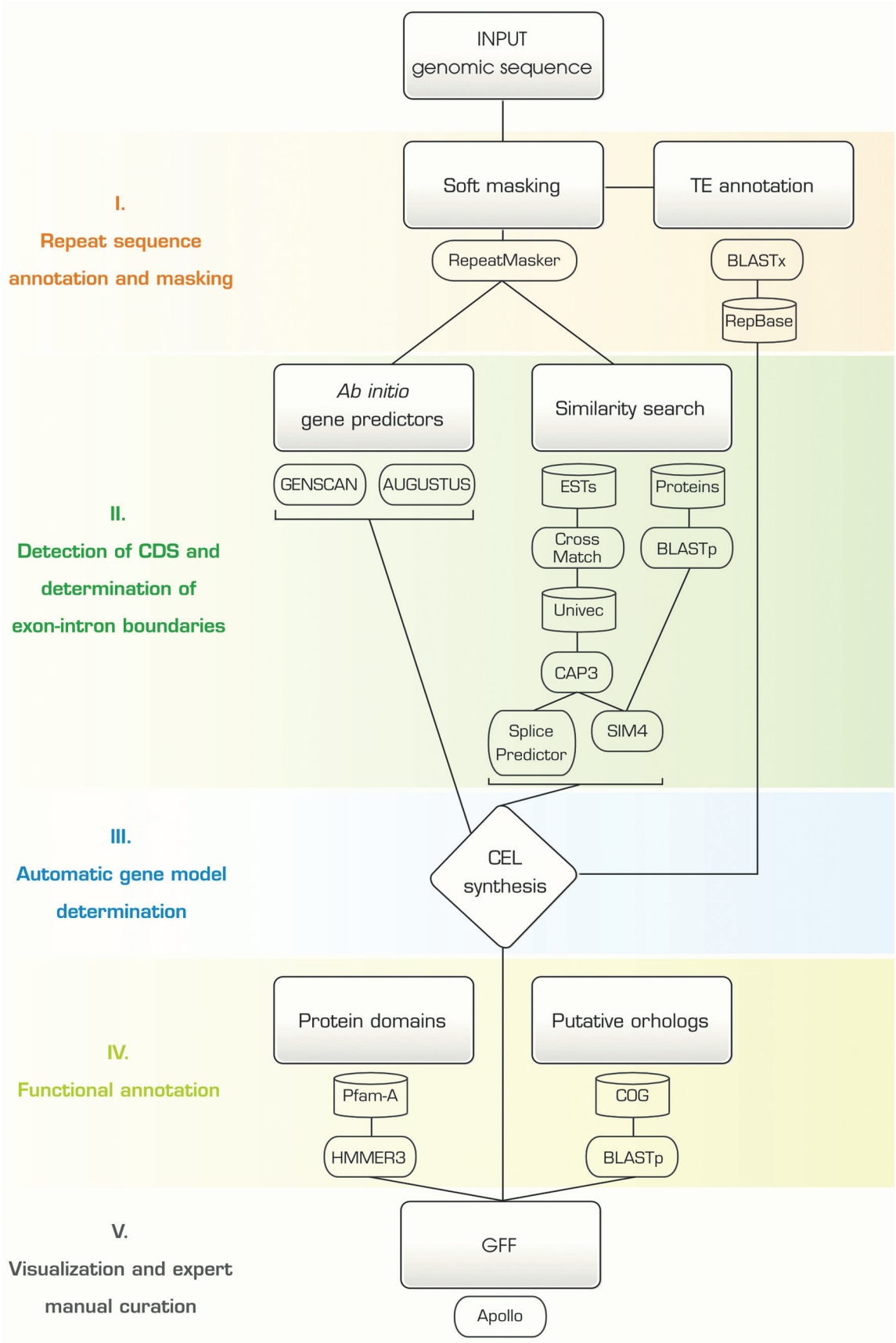

Fig. 1. An overview of the workflow supported by the CEL pipeline/CEL software architecture. Actions corresponding to the five basic steps of automatic annotation are shown in color panels. The detailed description of each step is provided in the text

in the EST relative to the genomic template. This means that non-cognate ESTs can be used for gene structure prediction, including ESTs derived from homologous genes from related species. This feature is thus very useful while annotating poorly-characterized genomes of non-model organisms.

III. Automatic gene model determination

The outputs of the computational analyses are used for automatic gene structure prediction. The CEL's algorithm incorporates biologists' expertise by combining precomputed diverse evidence from different methods to determine a gene model prediction whose intron-exon structure best represents the consensus of the models from the overlapping predictions. Such approach quite closely mirrors the expert curation process in which annotators review the evidence for each analysis and each gene in order to decide the structure of the final consensus gene model. The gene prediction accuracy is influenced by the types of evidence provided and directly associated with "confidence values" correlated with them. For example, the results of *ab initio* methods achieve lower cumulative scores than evidenced-based results. The qualitative influence of each of the methods is determined experimentally by an expert annotator before the prediction process. It is possible to produce several versions of annotation with the same computational results by changing the "confidence values" for prediction methods. The default values of confidence to each method implemented in the CEL pipeline are based on the reannotation of the *Arabidopsis thaliana* genome, and can be changed by the user. In the CEL pipeline, all the annotation results are represented by separate layers of evidence (i.e. multiple gene predictions as well as transcript alignments) which are divided into sets of nonredundant gene structure components: exons and introns. In the final prediction, each exon and intron achieves a score based on the weight (associated numerical confidence value) and abundance of the supporting evidence. The combination of highest scoring gene elements is then used to predict a gene structure. For example, an exon supported by multiple transcript/protein alignments will reach a higher score than an alternate exon of similar length supported by only a single similarly weighted transcript/protein alignment. Ultimately, the user is presented with a most representative gene structure as a weighted consensus of all available evidences. The overall quality of such model can be further evaluated by the user's inspection of details of any prediction layer. Therefore, the user has complete control on the final outcome of the annotation process.

IV. Functional annotation

For functional assignment predicted protein sequences are subjected to a series of metagenomic comparisons to existing, previously annotated sequence records. First step includes detection of the protein domain organization by searching PFAM, a collection of Hidden Markov Model (HMM) profiles and alignments for common protein families (Punta et al., 2012) with the HMMER 3.0 program (Finn et al., 2011). Next, putative orthologs (functionally equivalent homologous sequences in other genomes) are identified for each gene model by BLAST-based similarity-search of the COG (Cluster of Orthologous Groups of proteins) database (Tatusov et al., 2003). Finally, the Gene Ontology (GO) terms are directly assigned to each gene prediction based on the results from PFAM and COG searches. The GO is a standard vocabulary for describing the function of individual genes in the context of the cell. It consists of three divisions: molecular function, biological process and cellular component. The molecular function terms describe the tasks carried out by individual gene products, such as its enzymatic activity or structural function. The biological process terms are used for broader biological goals, such as meiosis or signaling cascade. The cellular component terms describe genes in terms of the subcellular structures they are localized to, such as organelles as well as the macromolecular complexes they belong to, such as the ribosome (Harris et al., 2004).

V. Visualization and expert manual curation

Once the computations are complete, the prediction results (outcomes of prediction methods and the automatic gene structure prediction) generated by CEL can be directly viewed in any genome annotation tool that supports GFF annotation format. In our study we used an open-source Apollo annotation editor which provides several powerful tools to verify and refine annotations manually (Lewis et al., 2002). In Apollo, the data supporting each annotation evidence and/or any genomic feature are shown as an independent layer on a segment of genomic DNA, which conforms well with the philosophy of data handling by the CEL annotation system.

In this way, gene models can be graphically revised using mouse's clicks and drag-and-drops, and the results can be saved to files or a remote database. In addition to results which can be visualized by GFF-aware genomic viewer, the CEL system provides a detailed log file with a summary of the entire annotation process.

De novo annotation of narrow-leafed lupin, non-model plant species

Since narrow-leafed lupin lacks extensive genomic resources, we collected a high-quality non-redundant sets of EST, mRNA and protein sequences for already completely sequenced legume genomes: *Medicago truncatula*, *Lotus japonicus* and soybean (*Glycine max*). We also used resources that have been recently developed to different degrees for other major grain and pasture legume crops, including pea (*Pisum sativum*), common bean (*Phaseolus vulgaris L.*), mung bean (*Vigna radiate*), chickpea (*Cicer arietinum*), cowpea (*Vigna unguiculata*), pigeon pea (*Cajanus cajan*), groundnut (*Arachis hypogaea*) and clover (*Trifolium repens*). Table 2 summarizes the number of sequence records and the main resources for each species in the legume family.

In order to demonstrate CEL's potential for application to emerging model organisms, we annotated BAC clones and BAC-end sequences of narrow-leafed lupine. Both, BAC libraries and BAC-end sequences are valuable resources which have contributed significantly to genetic and genomic studies of a wide range of models or economically important plant species (Varshney et al., 2010, Kasprzak et al., 2006). The CEL's annotation of five BAC clones (HE804808, HE804809, HE804810, HE804811, HE804812) and 210 BAC-end sequences retrieved from EMBL, totaling ~0.5 Mb of DNA, have led to the identification of 33 protein-coding genes, 14 retroelements and 3 DNA transposons. A summary of gene ontology has revealed a relatively higher proportion of proteins involved in polymerization, microtubule-based movement, regulation of transcription, metabolism and SRP-dependent cotranslational targeting to membrane (Fig. 2). To present an example outcome of the CEL's predictions, from among the five BACs, we selected the most gene abundant contig of ~45 Kbp and displayed its annotation in the Apollo genome browser (Fig. 3A). The computational evidence assembled by CEL on both forward (upper panel) and the reverse (bottom panel) strands are shown as color-coded layers in the black panels: Augustus (light blue), GENSCAN (green), EST assemblies (white), homologous mRNA (orange) and proteins (yellow). Evidence gathered by CEL's compute pipeline is combined into the resulting CEL annotation (light blue panels) of five genes with exon/intron structures are similar to their counterpart genes found in other plants (orange and yellow layers). Four gene models – 1, 2, 3, 4 – are predicted with highest confidence since their exons are supported by both similarity-based methods: homogeneous ESTs and mRNA/protein sequences form closely related legume species. Products of 1st and 4th gene models, though highly conserved among legume family, are annotated as proteins of unknown functions. The 2nd gene model is represented by transcript encoding MYB transcription factor which performs a variety of functions in developmental and stress response processes in plants (Zhang et al., 2012). The 3rd gene model is homologous to gene that encodes the isoflavone reductase (IFR), an enzyme involved in biosynthesis of isoflavonoid phytoalexins in legumes that plays an important role in plant defense and exhibits a range of mammalian health-promoting activities (Dixon and Steele, 1999). Although the 5th gene model is not covered by homogeneous EST evidence, probably due to the scanty resources of lupins, it is conserved among green plants. Polypeptide predicted from the gene sequence contains START and pleckstrin homology domains, both are involved in intracellular signaling or act as constituents of the cytoskeleton (Pfam accessions: PF01852 and PF00169, respectively).

Low- and high-confidence gene annotations

The CEL system is capable of accommodating a variety of evidence types, including (but not limited to) gene models computed by diverse gene finders, BLAST hits, EST matches, and splice site predictions. Therefore, it is possible to prioritize any analysis type according to the confidence in the underlying supporting evidence. This priority values can be easily changed by the user and assigned to any CEL component analysis. This feature is very useful while looking for a highly reliable subset of annotating genes. To demonstrate the impact of different confidence levels in annotation results we assumed simple scoring values (Augustus = 1, GENSCAN = 1, homologous mRNA and Proteins = 2 and

Table 2. Main resources used for narrow-leafed lupin genome annotation

Plant species	Number of records			
	mRNA	EST	GSS	Proteins
Arachis hypogaea	89 976[2]	178 490[2]	17 042[2]	864[1]
Cicer arietinum	1 194[2]	44 157[2]	51 511[2]	4 911
Glycine max	121 2992	1 461 624	368 5882	86 498[1,2,3]
Lotus japonicus	15 168[2]	242 432[2]	46 569[2]	8 675[3,4,5]
Medicago truncatula	68 660[2]	269 238[2]	168 929[2]	56 616[3,4,6]
Phaseolus vulgaris	4 455[1]	116 509[1,7]	92 237[1]	2 493[1]
Pisum sativum	102 365[1]	18 576[1]	204[1]	3 274[1]
Trifolium pratense	319[1]	38 109[1]	99 970[1]	222[1]
Vicia faba	6 834[1]	5 415[1]	529[1]	478[1]
Vigna unguiculata	606[2]	187 487[2]	54 949[2]	28 359[8]
Lupinus angustifolius	172[1]	388[1]	14 224[1]	115[1]
Cajanus cajan	65 440[2]	25 576[2]	90 108[2]	42[2]

[1] UniprotKB: http://www.uniprot.org/ [2] GenBank: http://www.ncbi.nlm.nih.gov/genbank/ [3] PlantGDB: http://www.plantgdb.org/ [4] TFDB Legume: http://legumetfdb.psc.riken.jp/ [5] MIPS: http://mips.helmholtz-muenchen.de/plant/lotus/ [6] Medcago http://medicago.org/ [7] KEGG: http://www.genome.jp/kegg/ [8] LIS: Legume Information System: http://www.comparative-legumes.org/

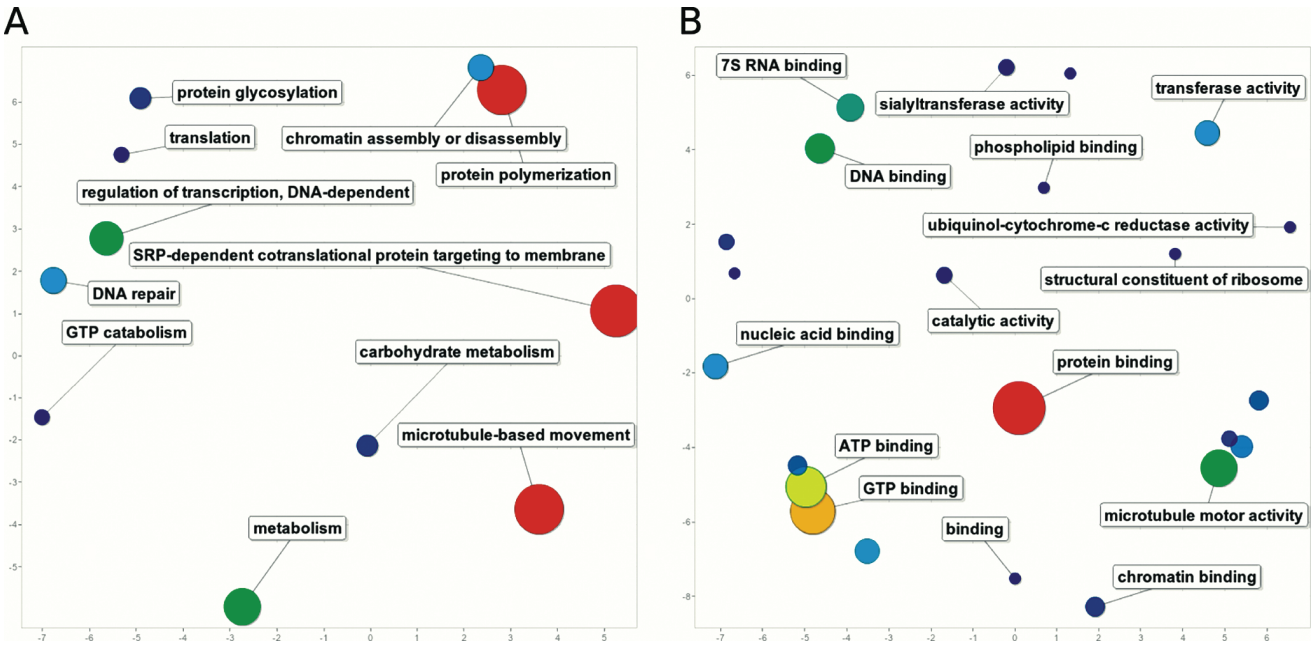

Fig. 2. Principal Component Analysis (PCA) scatter-plots (generated by REViGO) of the abundance of gene ontology (GO) terms related to **A.** biological processes and **B.** molecular function. GO terms are represented by circles. Circles representing similar GO terms are clustered closer together than circles representing unrelated GO terms. The sizes (big = high, small = low) and colors (red = high, green = moderate, blue = low) of circles are proportional to the numbers of functional annotations (GO terms) predicted in the five BAC clones of *Lupinus angustifolius*

homogeneous ESTs = 3) to annotate a gene that is ambiguously supported by different types of evidence. Using the example of a gene encoding the kinase family member predicted by CEL in BAC004G15 (EMBL accession: HE804808) in Figure 4B we have demonstrated various gene structure models (light blue panel) generated by CEL at different cumulative scoring thresholds: from 1 to 5. It can be easily observed that by applying

Fig. 3. Apollo graphical presentation of *de novo* annotation of *Lupinus angustifolius* genomic sequences: A) ~45-kbp sequence contig from BAC080B11 (EMBL Accession: HE804812). Curated annotations on both forward strand (upper panel) and reverse strand (bottom panel) are displayed in light blue panels. Evidence tiers are shown in the black panels: Augustus (light blue), GENSCAN (green), EST assemblies (white), orthologous mRNA (orange) and proteins (yellow); B) Annotation of the kinase gene from BAC004G15 (EMBL accession: HE804808) using confidence levels of different evidence and applying different thresholds

higher thresholds, CEL retains high-confidence exons. Although the resulting gene models of the highest thresholds do not represent all the elements of the exact gene structure, the presence of high-confidence exons in the model usually guarantees at least partially correct translations, which should make them useful in further genome analysis.

Reannotation of lupin asparate aminotransferase

The quality of annotating new genomes is strongly dependent on the availability of reference transcribed sequences. Such genomic projects generally lack pre-existing, gene-defining data, hence it highlights the need for efficient annotation pipelines that can use all usable information to extract functional information from new acquired sequences. One of our main goals was to study how well the CEL pipeline performs on genomes represented by limited, incomplete or exclusively cross-species genetic data. To benchmark the CEL performance we re-annotated selected *Lupinus* genes retrieved from NCBI. In Figure 4 we present the result of CEL's gene annotation of a single gene, aspartate aminotransferase P1 (AAT). AAT plays an important role in nitrogen metabolism in all plants and is particularly important in the assimilation process of fixed N during the legume-*Rhizoblum* symbiosis (Gantt et al., 1992). The presented annotation results include *ab initio* gene predictions (shown as a green layer) and heterogeneous types of evidence such as the collection of available ESTs from lupin, as well as homologous mRNAs and proteins from other plant species. Both *ab initio* programs, Augustus and GENSCAN, predict gene models consisting of 12 exons with slight differences in exon 3 and 12. Although, exons 2, 3 and 4 are not supported by ESTs, there is evidence of their presence covered by mRNA and protein sequences of aspartate aminotransferases in other plant species. The gene model determined by CEL, shown as a red layer, is a result of the unification of the evidence and prediction data from all layers. We directly compared this prediction with a reference gene model produced by a multi-step and computationally intensive NCBI annotation pipeline (Pruitt et al., 2012). The per-

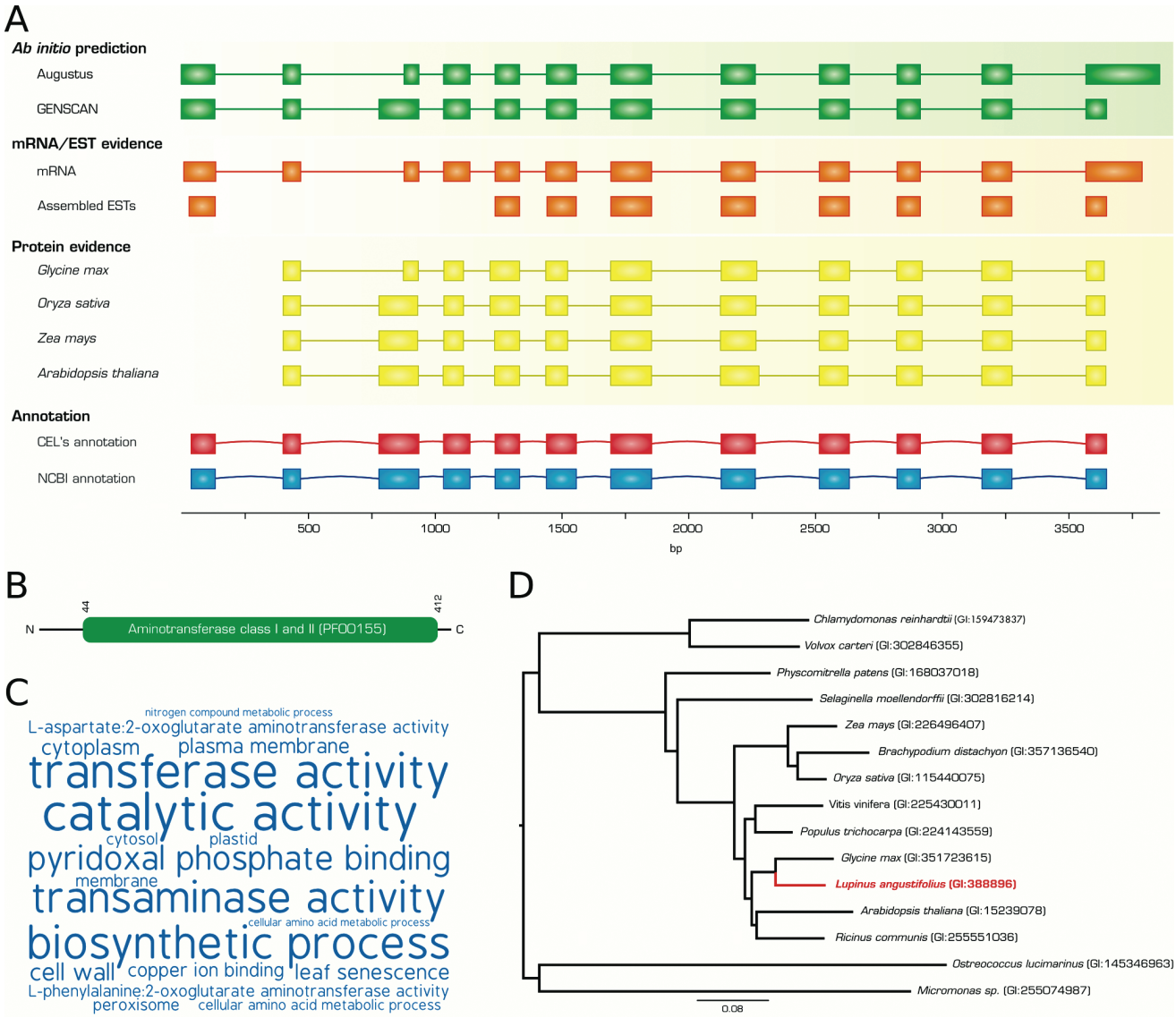

Fig. 4. Re-annotation of aspartare aminotransferase (AAT): A) Structural annotation of the AAT gene and its associated predictions and homology-based evidence. Green, orange and yellow layers show the *ab initio* predictions, EST/mRNA and proteins, respectively, that are produced during the computation phase of the annotation process. The evidence gathered by CEL's computational pipeline is synthesized into the resulting gene model (shown in red) that is identical to the reference NCBI model (shown in blue); B) Schematic representation of *Lupinus angustifolius* AAT protein, member of aminotransferase class I and II (PF00155) family; C) Data cloud representation of Gene Ontology (GO) descriptions of orthologous plant AAT proteins; D) Phylogeny of the AAT family in orthologous plant AAT proteins

fect agreement in gene structure between CEL and NCBI pipelines soundly confirms the quality of the prediction data coming from our system.

Functional annotation of the predicted AAT gene (comparative genomics between ATT gene and other plant species, see Annotation system implementation) includes: identification of potential protein domains, detection of putative orthologs and the GO term assignment. As expected, the predicted protein is identified in Pfam as a member of the aminotransferase family (Fig. 4B). Based on orthologous proteins found in all model plant genomes (Fig. 4D), GO terms are assigned to the predicted AAT gene (Fig. 4C) again suggesting the transaminase activity of the encoded protein.

Conslusions

The growing number of sequenced genomes calls for an easy-to-use, handy, and yet comprehensive and multi-

platform software for gene annotation. An analysis of sequences from non-model species requires special attention because of the lack of dedicated reference resources and specialized prediction methods. Here, we present a new annotation system which combines well-tested solutions from other projects with modular structure, intuitive configuration and unlimited expanding capabilities. As presented in the examples of NLL genomic DNA annotation, our solution is well designed for annotation of the sequences from species located at the frontiers of genomic research.

The CEL pipeline combines all necessary steps required during gene identification, structure prediction and functional annotation. It incorporates tools widely used during genome annotation, yet it does not require high-end computers and sophisticated operating systems to function. Additionally, our annotation system is easy to use and install, and poses unlimited expanding capabilities that allow incorporation of new, better suited for a particular project tools. In the current form, our pipeline was optimized to work well for lupine sequence and therefore is a first publicly available tool dedicated for this species and should obviously be of interest for research groups working on those plants.

Using well-defined sequences from the databank, we have proved the quality of our annotation system. Two other examples of annotation of lupine genomic DNA show the power of the CEL system which can accommodate even the most complex tasks. In our system, the user decides on the composition of the reference dataset. Additionally, all the partial and final results are presented in universal and widely-accepted formats which makes it possible to snap-preview the annotation process and export the results of analyses to any modern genomic feature viewer.

In conclusion, the annotation system presented here is a proof that a clear and simple design with carefully selected tools and open architecture represents a right strategy for the development of computational tools for genome analyses and general biocomputing applications.

Availability and requirements

The software is freely available for non-profit users upon request. It requires a Python interpreter version greater than 2.6. Default CEL's components, described in the text, are integrated in the package.

Acknowledgements

The financial support was provieded by the Faculty of Biology, Adam Mickiewicz University to A.Z., P.P. and W.M.K. M.K was supported by two Polish Ministry of Science and Higher Education research grants: PBZ-MNiSW-2/3/2006/3 and N N301 391939.

References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acids Res. 25: 3389-3402.

Brendel V., Xing L.Z.W. (2004) *Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus.* Bioinformatics (Oxford, England), 20(7): 1157-1169.

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

Cantu D., Vanzetti L.S., Sumner A., Dubcovsky M., Matvienko M., Distelfeld A., Michelmore R.W. et al. (2010) *Small RNAs, DNA methylation and transposable elements in wheat.* BMC Genom. 11: 408.

Carver T., Harris S.R., Berriman M., Parkhill J., McQuillan J.A. (2012) *Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.* Bioinformatics (Oxford, England) 28(4): 464-469.

de Koning A.P.J., Gu W., Castoe, T.A., Batzer M.A., Pollock D.D. (2011) *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet. 7(12): e1002384.

Dixon R., Steele C. (1999) *Flavonoids and isoflavonoids – a gold mine for metabolic engineering.* Trends Plant Sci. 4(10): 394-400.

Donlin M.J. (2007) *Using the Generic Genome Browser (GBrowse). Current protocols in bioinformatics*, ed. A.D. Baxevanis et al., chapt. 9, unit 9.9.

Duranti M., Consonni A., Magni C., Sessa F., Scarafoni A. (2008) *The major proteins of lupin seed: characterisation and molecular properties for use as functional and nutraceutical ingredients.* Trends Food Sci. Technol. 19: 624-633.

Ebert M.S., Sharp P.A. (2012) *Roles for microRNAs in conferring robustness to biological processes.* Cell 149(3): 515-524.

Erba M., Certel M., Uslu M.K. (2005) *Some chemical properties of white lupin seeds (Lupinus albus L.).* Food Chem. 89: 341-345.

Finn R.D., Clements J., Eddy S.R. (2011) *HMMER web server: interactive sequence similarity searching.* Nucl. Acids Res. 39: W29-37.

Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P. et al. (2012) *Ensembl 2012.* Nucl. Acids Res. 40: D84-90.

Florea L., Hartzell G., Zhang Z., Rubin G.M., Miller W. (1998) *A computer program for aligning a cDNA sequence with a genomic DNA sequence*. Genome Res. 8(9): 967-974.

Gantt J.S., Larson R.J., Farnham M.W., Pathirana S.M., Miller S.S., Vance C.P. (1992) *Aspartate aminotransferase in effective and ineffective alfalfa nodules: cloning of a cDNA and determination of enzyme activity, protein, and mRNA levels.* Plant Physiol. 98(3): 868-878.

Gao L.L., Hane J.K., Kamphuis L.G., Foley R., Shi B.J., Atkins C.A., Singh K.B. (2011) *Development of genomic resources for the narrow-leafed lupin (Lupinus angustifolius): construction of a bacterial artificial chromosome (BAC) library and BAC-end sequencing.* BMC Genomics 12(1): 521.

Harris M.A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., Eilbeck K. et al. (2004) *The Gene Ontology (GO) database and informatics resource.* Nucl. Acids Res. 32: D258-261.

Huang X., Madan A. (1999) *CAP3: A DNA sequence assembly program.* Genome Res. 9(9): 868-877.

Kasprzak A., Safár J., Janda J., Dolezel J., Wolko B., Naganowska B. (2006) *The bacterial artificial chromosome (BAC) library of the narrow-leafed lupin (Lupinus angustifolius L.).* Cell. Mol. Biol. Lett. 11(3): 396-407.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. (2002) *The Human Genome Browser at UCSC.* Genome Res. 12(6): 996-1006.

Kersey P.J., Lawson D., Birney E., Derwent P.S., Haimel M., Herrero J., Keenan S. et al. (2010) *Ensembl Genomes: extending Ensembl across the taxonomic space.* Nucl. Acids Res. 38: D563-569.

Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O., Walichiewicz J. (2005) *Repbase Update, a database of eukaryotic repetitive elements.* Cytogen. Genome Res. 110(1-4): 462-467.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K. et al. (2001) *Initial sequencing and analysis of the human genome.* Nature 409(6822): 860-921.

Lee Y.P., Mori T.A., Sipsas S., Barden A., Puddey I.B., Burke V., Hall R.S. et al. (2006) *Lupin-enriched bread increases satiety and reduces energy intake acutely.* Amer. J. Clin. Nutr. 84(5): 975-980.

Lewis S.E., Searle S.M.J., Harris N., Gibson M., Lyer V., Richter J., Wiel C. et al. (2002) *Apollo: a sequence annotation editor.* Genome Biol. 3(12).

Liang C., Mao L., Ware D., Stein L. (2009) *Evidence-based gene predictions in plant genomes.* Genome Res. 19(10): 1912-1923.

Lipman D.J., Pearson W.R. (1985) *Rapid and sensitive protein similarity searches.* Science (New York, N.Y.), 227(4693): 1435-1441.

Mathé C., Sagot M.F., Schiex T., Rouzé P. (2002) *Current methods of gene prediction, their strengths and weaknesses.* Nucl. Acids Res. 30(19): 4103-4117.

Makalowska I., Lin C.F., Makalowski W. (2005) *Overlapping genes in vertebrate genomes.* Comput. Biol. Chem. 29(1): 1-12.

Marquez Y., Brown J.W., Simpson C.G., Barta A., Kalyna M. (2012) *Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis.* Genome Res. 22(6): 1184-1195.

Nelson M.N., Moolhuijzen P.M., Boersma J.G., Chudy M., Lesniewska K., Bellgard M., Oliver R.P. et al. (2010) *Aligning a new reference genetic map of Lupinus angustifolius with the genome sequence of the model legume, Lotus japonicus*. DNA Res. 17(2): 73-83.

Pagani I., Liolios K., Jansson J., Chen I.M., Smirnova T., Nosrat B., Markowitz V.M. et al. (2012) *The Genomes On-Line Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.* Nucl. Acids Res. 40: D571-579.

Pearson W.R., Lipman D.J. (1988) *Improved tools for biological sequence comparison*. Proc. natl. Acad Sci USA 85(8): 2444-2448.

Pruitt K.D., Tatusova T., Brown G.R., Maglott D.R. (2012) *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucl. Acids Res. 40: D130-135.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N. et al. (2012) *The Pfam protein families database*. Nucl. Acids Res. 40: D290-301.

Saha S., Bridges S., Magbanua Z.V., Peterson D.G. (2008) *Empirical comparison of ab initio repeat finding programs.* Nucl. Acids Res. 36(7): 2284-2294.

Slater G.S.C., Birney E. (2005) *Automated generation of heuristics for biological sequence comparison*. BMC Bioinformat. 6: 31.

Sleator R.D. (2010) *An overview of the current status of eukaryote gene prediction strategies*. Gene 461(1-2): 1-4.

Smith T.F., Waterman M.S. (1981) *Identification of common molecular subsequences*. J. Mol. Biol. 147(1): 195-197.

Smith C.D., Edgar R.C., Yandell M.D., Smith D.R., Celniker S.E., Myers E.W., Karpen G.H. (2007) *Improved repeat identification and masking in Dipterans.* Gene 389(1): 1-9.

Stanke M., Morgenstern B. (2005) *AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.* Nucl. Acids Res. 33: W465-467.

Stein L. (2001) *Genome annotation: from sequence to biology.* Nature Rev. Genet. 2(7): 493-503.

Supek F., Bošnjak M., Škunca N., Šmuc T. (2011) *REVIGO summarizes and visualizes long lists of gene ontology terms*. PloS One 6(7): e21800.

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M. et al. (2003) *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics 4: 41.

The Arabidopsis Information Resource (TAIR), www.arabidopsis.org/aboutarabidopsis.html, on www.arabidopsis.org, Jun 27, 2012

Quesada V., Ponce M.R., Micol J.L. (1999) *OTC and AUL1, two convergent and overlapping genes in the nuclear genome of Arabidopsis thaliana*. FEBS Lett. 461(1-2): 101-106.

Yandell M., Ence D. (2012) *A beginner's guide to eukaryotic genome annotation.* Nature Rev. Genet. 13(5): 329-342. Nature Publishing Group.

Yok N.G., Rosen G.L. (2011) *Combining gene prediction methods to improve metagenomic gene annotation.* BMC Bioinformatics 12(1): 20. BioMed Central Ltd. doi: 10.1186/1471-2105-12-20.

Varshney R.K., Glaszmann J.C., Leung H., Ribaut J.M. (2010) *More genomic resources for less-studied crops.* Trends Biotechnol. 28(9): 452-460.

Zhang L., Zhao G., Xia C., Jia J., Liu X., Kong X. (2012) *Overexpression of a wheat MYB transcription factor gene, TaMYB56-B, enhances tolerances to freezing and salt stresses in transgenic Arabidopsis*. Gene 505(1):100-107.