BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 9 2007, pages 1061–1067
doi:10.1093/bioinformatics/btm071

Genome analysis

CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes

Genis Parra[1], Keith Bradnam[1] and Ian Korf[1,2,*]

[1]UC Davis Genome Center, 451 E. Health Sciences Drive and [2]Department of Molecular and Cellular Biology, University of California Davis, Davis, CA 95616, USA

Received on December 7, 2006; revised on January 26, 2007; accepted on February 22, 2007

Advance Access publication March 1, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The numbers of finished and ongoing genome projects are increasing at a rapid rate, and providing the catalog of genes for these new genomes is a key challenge. Obtaining a set of well-characterized genes is a basic requirement in the initial steps of any genome annotation process. An accurate set of genes is needed in order to learn about species-specific properties, to train gene-finding programs, and to validate automatic predictions. Unfortunately, many new genome projects lack comprehensive experimental data to derive a reliable initial set of genes.

Results: In this study, we report a computational method, CEGMA (Core Eukaryotic Genes Mapping Approach), for building a highly reliable set of gene annotations in the absence of experimental data. We define a set of conserved protein families that occur in a wide range of eukaryotes, and present a mapping procedure that accurately identifies their exon–intron structures in a novel genomic sequence. CEGMA includes the use of profile-hidden Markov models to ensure the reliability of the gene structures. Our procedure allows one to build an initial set of reliable gene annotations in potentially any eukaryotic genome, even those in draft stages.

Availability: Software and data sets are available online at http://korflab.ucdavis.edu/Datasets.

Contact: ifkorf@ucdavis.edu

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

1 INTRODUCTION

The pace of genome sequencing continues to increase and these new genomes contain a wealth of information that will be studied for years to come. The first question asked of a newly sequenced genome is usually 'how many genes does it contain'? This is generally followed by 'how many genes are unique to the organism'? Unfortunately, accurately annotating genes in eukaryotic genomes is a difficult task. Even in the best-case scenario where a genome project has a large body of experimental data and employs dedicated expert biologists to annotate gene structures, gene catalogs are still unfinished and under constant curation. The recent EGASP experiment (Harrow *et al*., 2006) shows that (1) computational methods are still less accurate than experts, and (2) even where experimental data is plentiful, some novel genes can still be predicted and verified. The situation is much worse for emerging genome projects, because there may be little or no experimental data. Only with the help of computational methods can we try to accomplish the challenging task of annotating all of the genomes that are going to be sequenced. So far, there is no affordable experimental system to provide annotations for new genomes. We have to rely on computational tools to generate, at least, the initial set of annotations. The rapid release of completed genome sequences has led to significant developments in genome annotation and gene finding tools.

Computational gene finding methods can be loosely categorized as being alignment-based, composition-based or a combination of both. Alignment-based methods can be used when trying to predict a gene that encodes a protein for which a closely related homolog exists. The DNA sequence of the gene is aligned to the protein or cDNA sequence of the homolog; gaps in the resulting alignment are presumed to correspond to potential introns in the gene (as long as they are compatible with known splicing signals). This is the approach in GeneWise (Birney *et al*., 2004) and PROCRUSTES (Gelfand *et al*., 1996). Composition-based algorithms (also known as *ab initio* gene-finding methods) contain a probabilistic model of gene structure based on biological signals (splice sites and translational start/stop sites) and compositional properties of functional sequences (coding, intron and intergenic). Unlike alignment-based methods, these algorithms rely only on the intrinsic properties of genes in order to build predicted gene structures. Genscan (Burge and Karlin, 1997) and geneid (Parra *et al*., 2000) are the two examples of this approach and they can find both known genes and novel genes as long as the genes fit the underlying probabilistic model. Combinations of both gene-finding methods have been developed where the results of searching a query sequence against a database of known coding sequences are then incorporated into the scoring schema of an *ab initio* gene-prediction method. The GenomeScan (Yeh *et al*., 2001) program is an example of this strategy as it incorporates protein to DNA alignments generated by the similarity search program BLASTX (Altschul *et al*., 1990) into gene predictions made by the Genscan program.

*To whom correspondence should be addressed.

Recent developments that further exploit sequence similarity come from comparative gene prediction programs. Instead of comparing an anonymous genomic sequence to known coding sequences, a genomic sequence is compared to anonymous genomic sequences from different species. This approach assumes that matching regions of conserved sequence will tend to correspond to coding exons. The SGP2 (Parra *et al*., 2003) and TWINSCAN (Korf *et al*., 2001) programs are examples of this strategy (see Brent, 2005 for a review).

Ensembl (Curwen *et al*., 2004) and other genome annotation pipelines attempt to take all sources of information into account. Their goal is to replicate expert biologists at genome centers. While these pipelines are largely evidence-based, *ab initio* gene prediction plays a major role in minimizing the search space and/or providing gene structures in the absence of evidence. *Ab initio* gene finders must be trained prior to their use in a particular genome. Training sets are usually derived from full-length cDNAs, but as previously mentioned, emerging genome projects may not have any experimental transcript data. Obtaining this set of cDNAs can be tedious and expensive. As a consequence, many new genome projects use a gene finder trained for some other genome. This is unfortunate because gene prediction is sensitive to genome-specific parameters, and one must train gene finders for individual genomes for optimal accuracy (Korf, 2004). With the amount of genome projects that are underway, there is a pressing need for an automated way of producing a reliable set of genes for each genome.

The availability of many complete genome sequences allows for the construction of evolutionary relationships between the genes that they encode. Orthologous genes are most likely to have maintained sequence conservation over evolutionary time, reflecting their conserved function. Classifying genes based on orthologous relationships appears to be a natural framework for comparative genomics and should facilitate the functional annotation of genomes. Thus, when comparing genes from two different genomes, the orthologs are likely to be those pairs of genes whose proteins exhibit the greatest sequence similarity. The Cluster of Orthologous Groups (COGs or KOGs for eukaryotic genomes) database (Tatusov *et al*., 2003) follows this approach and contains protein families (orthologs) of genes from a set of diverse species. Some of these genes are involved in fundamental biological pathways, and therefore, the degree of conservation they exhibit is higher than other less constrained proteins.

In this article, we introduce a computational method to obtain a set of reliable gene structures in any eukaryotic genome. Our goal is not to provide the complete catalog of genes in a genome, but to generate a highly accurate set of genes for those genomes without experimental data. The strategy relies on a simple fact: some highly conserved proteins are encoded in essentially all eukaryotic genomes. We use the KOGs database to build a set of these highly conserved, ubiquitous proteins. We call these protein families *core proteins* and the genes that encode them *core genes*. We define a set of 458 core proteins and present an accurate mapping protocol that maps the likely ortholog of each gene in a genomic sequence and then predicts the exon–intron structure. We show that our procedure does not need any previous knowledge of the target genome, that it is highly accurate, and that it can be used even for those genomes in draft stages.

An approach based on similar principles has already been described in (Natale *et al*., 2000) to find undetected proteins in prokaryotic genomes using the COGs database. Due to the lack of introns in prokaryotic genes this approach is more simplistic, relying chiefly on BLAST searches. The more complex gene structures of eukaryotes limits the use of this approach in non-prokaryotic genomes. Our approach overcomes these limitations by combining BLAST searches, GeneWise and geneid to accurately detect complex eukaryotic gene structures. A key feature of the system is that it includes the use of protein profile-hidden Markov models to ensure the reliability of the predicted gene structures. To check the quality of the final annotations, predicted proteins are aligned against a profile derived from the corresponding core protein family. Only predictions that match the profile of the gene family are selected. This filtering protocol ensures a high accuracy of the mapping process.

2 METHODS

2.1 A set of eukaryotic core proteins

The eukaryotic orthologous groups (KOGs) database (Tatusov *et al*., 2003) was used to produce a set of core genes. The KOGs database contains groups of genes from the following species: *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. From the complete set of 4852 groups (also called KOGs) we selected 1788 that had at least one protein from each of the six species. A global multiple protein alignment was produced for each KOG using T-coffee (Notredame *et al*., 2000). As some of the KOGs contained more than one protein for each species, the information given by the T-coffee alignment was used to select the protein of each organism most similar to the global alignment. After that, the selected protein sequences of each species were aligned again with T-coffee to generate the final multiple alignments. We surveyed the alignments and found that many contained extended gaps and misaligned regions. Some possible explanations could be the partial or incomplete annotations of the proteins, the misclassification of paralogous genes instead of the real ortholog, or the presence of proteins which are more evolutionary divergent in certain species. The following criteria were used to remove dubious alignments: (1) all proteins must cover at least 75% of the length of the global alignment, (2) no more than five internal gaps longer than 10 amino acids for each aligned protein were allowed and (3) the average percent identity over all rows in the alignment must be >10%. Use of these criteria reduced the data set to 458 KOGs. Figure 1 shows the flowchart of this process.

2.2 Genomes

The versions of the genome sequences that were used are as follows: *A.gambiae* Feb 2003, *A.thaliana* R5v01212004, *C.elegans* WS140, *C.reinhardtii* v.3.0, *C.intestinalis* 1.95, *D.melanogaster* 4.1, *H.sapiens* NCBI35 Ensembl build Nov 2004, *S.cerevisiae* July 06 2005, *S.pombe* July 04 2005, and *T.gondii* Tg10x 31-Draft3.

2.3 Sequence analysis

The measures of gene prediction accuracy that we used have been previously described (Burset and Guigo, 1996), and we first measured accuracy at the levels of nucleotides, signals (donor, acceptor, initiation

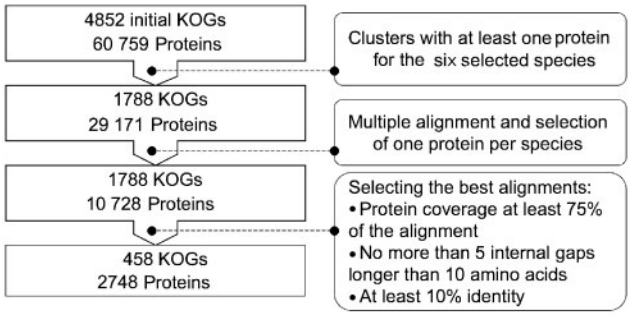

Fig. 1. Flow chart of the KOGs filtering protocol showing steps that filter the original set of KOG proteins to produce the final set of 458 core set of proteins.

and termination sites) and internal exons. We define *sensitivity* as the percentage of actual coding nucleotides/signals/exons that have been correctly predicted, and *specificity* as the proportion of the predicted coding nucleotides/signals/exons that are actually coding. To compute these measures at the exon level, we will only score an exon as correctly predicted when both its boundaries have been correctly predicted. At the exon level, we also compute statistics for the percentage of *missed exons* and *wrong exons*. Missed exons are real exons that do not overlap any predicted exons, and wrong exons are predicted exons that do not overlap any real exons. Finally, we calculate accuracy at the CDS level, where genes are correctly predicted if all of their exons are correct when compared to the known gene structure. Note that only a single accuracy value is provided for initiation/termination signals and CDSs as only one start, stop and CDS is predicted for each putative ortholog. The software used for the calculations is the *fathom* program distributed with the SNAP package (Korf, 2004).

2.4 Mapping protocol

We have developed a procedure, CEGMA (Core Eukaryotic Genes Mapping Approach) to find orthologs of core proteins in new genomes and to determine their exon–intron structures. A local version of CEGMA can be installed on UNIX platforms and it requires pre-installation of PERL, WU-BLAST (http://blast.wustl.edu), HMMER (http://hmmer.janelia.org), GeneWise (Birney *et al.*, 2004) and geneid (Parra *et al.*, 2003). The procedure uses information from the core genes of six model organisms by first using TBLASTN to identify candidate regions in a new genome. It then proposes and refines gene structures using a combination of GeneWise, HMMER and geneid. The system includes the use of a profile for each core protein family to ensure the reliability of the gene structures. Ultimately, in any new genome we attempt to predict gene structures for the orthologs of each of the 458 core proteins. The general schema of CEGMA is illustrated in Figure 2. Before testing the protocol on new genome sequences, we first test on each of the model organisms from which the core genes are derived. In doing so, no information is used from the species being tested, i.e. we use information from only five species to predict genes in the sixth. The following sections describe each step in detail.

2.5 Finding the location of core protein orthologs

For each core protein family, the alignment program TBLASTN is used to search the six proteins in each family against a new genome sequence. To speed up the process the word size parameter is set to 5 (W = 5) and the neighborhood word threshold score is set to 25 (T = 25). This step produces a number of candidate regions that might contain the ortholog of the core protein, though only the five best candidate regions

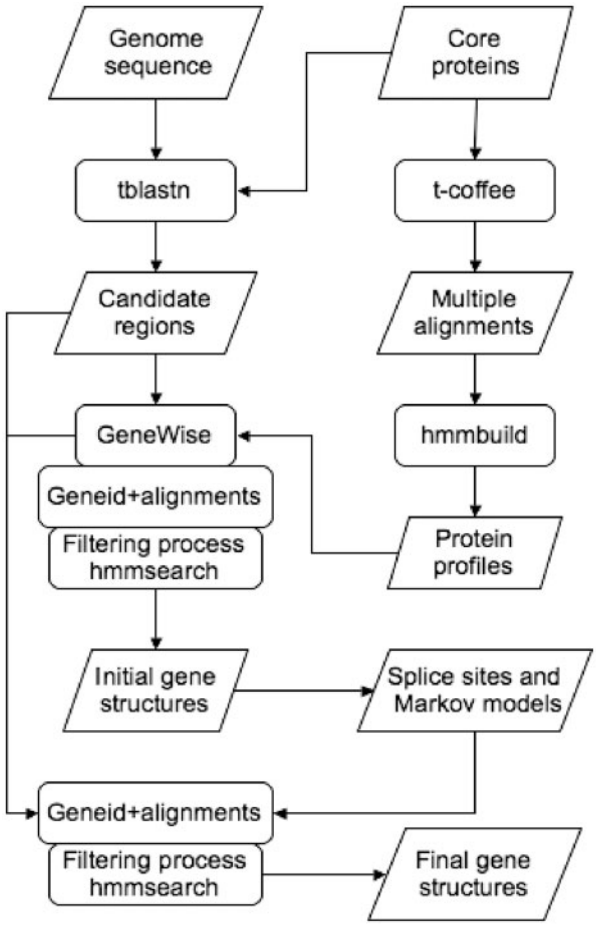

Fig. 2. Flowchart of CEGMA. The initial sources of information are the raw genomic sequence and the multiple alignment of the set of core proteins.

are considered further (B = 5 and V = 5). For the human genome, the sequence was split in fragments of 1 Mb with 100 Kb of overlapping sequence. To build the initial candidate regions, high scoring pairs (HSPs) closer than 5 Kb are clustered into a single candidate region. For each of the top five candidate regions, 2 Kb of flanking sequence is also extracted; the sequence is reverse complemented if the BLAST alignment shows similarity to the reverse strand. For the human genome the permitted distance between HSPs was increased to 40 Kb and the length of extracted flanking sequences was extended to 5 Kb.

2.6 Protein profile alignment

The candidate regions produced by TBLASTN are processed by GeneWise using a profile hidden Markov model that is built for each KOG multiple alignment [using the hmmbuild program (Eddy, 1998)]. Use of GeneWise increases sensitivity and specificity when compared to only using BLAST alone. For *C.elegans*, it is able to predict 96.6% of the coding regions compared to 89.5% from the translated BLAST approach (Table 1). GeneWise is a powerful tool for aligning protein profiles on to a genome, but as it lacks an inherent model of gene structure it is inappropriate for predicting complete gene structures. For instance, GeneWise does not have a model for the initial transcription site, and does not force the prediction to finish with a stop codon. The accuracy of the initiation and termination translational sites predicted

Table 1. Accuracy of the different steps of the CEGMA pipeline

Mapping step	Nucleotide	Internal exons	Donor sites	Acceptor sites	Missed exons	Wrong exons	Start sites	Stop sites	CDS
BLAST	89.5/84.7	–	–	–	15.1	30.3	–	–	–
GeneWise	96.6/93.4	82.4/70.1	88.2/77.2	90.0/78.8	4.6	13.8	26.0	56.1	12.9
GeneWise + geneid	97.5/97.6	88.6/86.7	91.4/89.9	93.4/92.0	4.1	5.9	71.9	91.0	57.5
GeneWise + geneid + self-training	98.3/96.3	91.6/90.0	93.4/91.4	95.1/93.1	3.2	5.3	78.1	92.8	64.5

CEGMA was run on the *C.elegans* genome using the core proteins of *A.thaliana*, *D.melanogaster*, *H.sapiens*, *S.cerevisiae* and *S.pombe*. No *C.elegans* information was used in any of the steps. BLAST refers to use of TBLASTN searches with the proteins of the five previously mentioned species. GeneWise corresponds to the accuracy of the profile of the orthologous proteins in the candidate regions. GeneWise + geneid corresponds to the initial geneid predictions based on consensus splice signals and GeneWise alignments. GeneWise + geneid + self-training refers to the second geneid predictions using species-specific coding statistics and gene signals estimated from the first set of annotations plus GeneWise alignments. The accuracy is measured from the final set of 442 mapped *C.elegans* genes (2204 exons). Values shown refer to sensitivity and specificity (Sn/Sp), all numbers are percentages.

by GeneWise is very low (26% and 56%, respectively). Furthermore, when aligning proteins or profiles from distantly related species, GeneWise can erroneously extend some of the alignments producing artefactual exons. This effect is depicted by the low specificity at exon level (70.1%) and the high percentage of wrong exons (13.8%, Table 1).

2.7 Refining the predicted gene structures

In order to improve the accuracy of the predicted gene structures, we produced a more complex gene-building strategy. This refined strategy attempts to extend homology results and remove spurious alignments so that even partially correct alignments can produce accurate gene structures. The geneid program is a gene finder that predicts and scores all potential exons within a specified sequence. Scores of exons are computed as log-likelihood ratios, which are a function of the splice sites defining the exon and of the compositional coding bias in the exon sequences [as measured by a Markov Model (Borodovsky and McIninch, 1993)]. From the set of predicted exons, geneid assembles a final gene structure, maximizing the sum of the scores of the assembled exons, using a dynamic programming chaining algorithm. This strategy has already been successfully used to integrate homology information in the SGP2 framework (Parra *et al.*, 2003). In our approach, geneid uses GeneWise alignments (instead of TBLASTX) to improve the scores of the *ab initio* predicted exons. The GeneWise alignment is converted to General Feature Format (GFF) for use by geneid. A detailed description of how such external information is integrated on geneid can be found at (Parra *et al.*, 2003).

For the initial geneid predictions, no coding model was used and geneid predicted exons using only the information from available start, stop and splice sites (weight matrices for splice sites and start sites were averaged from a mixture of all six species). The resulting gene structures were, therefore, largely driven by GeneWise predictions. We constrained geneid to predict a single, positive-strand gene by using options F and W. The combined use of geneid and GeneWise in this way produces more accurate predictions in *C.elegans* than when using GeneWise alone (Table 1). The accuracy at internal exon level increases by 10% and the amount of wrong exons decreases by 8%. More strikingly, there is a dramatic increase in the accuracy at initiation and termination sites (71.9% and 91%, respectively).

2.8 Verifying candidate proteins

After the initial round of geneid predictions, a filter was applied to the resulting gene structures to determine if they were similar enough to the rest of the KOG group to be considered orthologs. This step requires use of pre-calculated 'intra-KOG-similarity' scores. For each KOG, we align each component protein sequence to a profile generated from the remaining five sequences [using *hmmsearch* (Eddy, 1998)]. From this, we obtain six scores which are then averaged to produce an approximate indicator of similarity between the proteins in each KOG. We then take the protein sequences of the orthologs predicted by geneid and align to the KOG profiles. The geneid prediction is retained if the score of this match is greater than half of the intra-KOG-similarity score, otherwise the prediction is rejected. As up to five candidate genomic regions are considered for each ortholog, it is possible for more than one predicted protein to have a score above the cutoff. In these cases the protein with the highest scoring alignment to the profile is selected as the 'true' ortholog, and the remainder are considered as putative paralogs.

2.9 Self-training and final set of annotations

Whilst the results from using GeneWise and geneid are promising, we attempted to further improve the accuracy of the resulting gene predictions. As previously mentioned, geneid does not use a coding model to make its initial set of predictions. Accuracy can be improved if geneid is allowed to build a coding model by training from this initial set of predictions. We, therefore, take the gene predictions made by geneid and then train from them to produce species-specific coding (Markov chain of order 5) and splice site models (position weight matrices). With these species-specific coding models we then use geneid (again in combination with GeneWise) to re-predict the set of genes. This step includes re-predicting gene structures that previously were below the intra-KOG-similarity threshold. The final gene structures predicted through this self-training approach show further increases in accuracy. Although there is no improvement at the nucleotide level, accuracy is increased by 3% for internal exons level and by more than 10% for complete CDS predictions (last row, Table 1). This method (GeneWise plus geneid with self-training) was the chosen strategy for subsequent testing.

3 RESULTS

3.1 Identifying core proteins

We have collected a set of 458 core proteins that exist in a wide range of organisms from plants, to fungi, to mammals. An example of a typical core protein is shown in Figure 3. Most core proteins appear to encode house-keeping genes, with a wide range of different functions (Supplementary #1). Gene finding is generally a hard problem, but core proteins

Fig. 3. Multiple alignment of a typical core protein family. This family corresponds to the SAR1 GTPase involved in vesicle transport.

Table 2. Accuracy of the CEGMA pipeline

Genome	Mapped (exons)	Nucleotide	Internal	Donor	Acceptor	Missed	Wrong	Start	Stop	CDS
A.thaliana	440 (2333)	97.2/98.2	90.0/93.0	93.1/94.9	94.0/95.9	3.6	1.9	76.7	91.9	58.5
C.elegans	442 (2204)	98.3/96.3	91.6/90.0	93.4/91.4	95.1/93.1	3.2	5.3	78.1	92.8	64.5
D.melanogaster	456 (1543)	99.0/98.3	90.6/88.1	92.6/91.1	94.6/93.2	4.4	5.3	79.7	96.3	72.7
H.sapiens	451 (4354)	96.2/96.5	92.0/90.3	93.0/91.4	94.7/93.0	4.0	5.5	62.0	82.5	35.5
S.cerevisiae	427 (469)	99.8/99.6	–	91.3/72.4	91.3/72.4	0.8	3.3	91.5	100	91.0
S.pombe	449 (966)	99.6/98.5	82.4/88.8	88.8/93.5	88.0/92.7	3.4	2.3	96.9	96.9	80.4

The number of core genes found by the mapping procedure (Mapped) is out of a maximum of 458. The number of exons is showed in parentheses. The accuracy measures are computed using the genomic annotations as reference. Values shown refer to sensitivity and specificity (Sn/Sp), all numbers are percentages.

are so highly conserved that they can be found with sequence alignment programs like BLAST. Determining the exact exon-intron structure for the core proteins is still a difficult problem (see subsequently), but unlike typical gene finding, there is a control: we know that ultimately the encoded protein should resemble the rest of the family.

3.2 Mapping core proteins into the test genomes

To determine how well the mapping protocol performed, we first mapped core proteins in the genomes of *A.thaliana*, *C.elegans*, *D.melanogaster*, *H.sapiens*, *S.cerevisiae* and *S.pombe*. For these experiments, we left out the core proteins for the genome under investigation. That is, when evaluating the procedure on *A.thaliana*, for example, we did not include any *A.thaliana* proteins or profile-hidden Markov models in the mapping procedure. This allowed us to determine the accuracy of the mapping procedure in *A.thaliana* as if it was a new genome. We find that the mapping procedure finds virtually all orthologs and discriminates coding from non-coding nucleotides with >97% accuracy (Table 2, also see Supplement #2). It also finds >90% of the acceptor, donor and stop sites. The N-terminal regions of core proteins tend to be less highly conserved, and this may be why the procedure is slightly less accurate for start sites. The most accurate results were achieved for the *S.cerevisiae* genome, reflecting the lack of introns in most yeast genes.

An investigation of the core genes that were not mapped showed that there are several reasons why some genes were missed. The most common reason is when an ortholog of a core protein does get mapped to the correct region, but the resulting gene prediction omits one or more exons. When exons are skipped in this way, the mapped protein becomes too dissimilar when compared to the rest of the family, and is discarded. Exon skipping can happen when some exons (typically shorter ones) are less conserved than the rest of the gene and are therefore not always detected by BLAST.

3.3 Mapping core proteins to recently sequenced genomes

To determine how well the complete mapping and training procedure performs, we evaluated the mapping procedure on the recently sequenced genomes of *Ciona intestinalis* (Dehal *et al*., 2002), *Toxoplasma gondii* (Kissinger *et al*., 2003), *Anopheles gambiae* (Holt *et al*., 2002) and *Chlamydomonas reinhardtii* (Grossman *et al*., 2003). These genomes were sequenced using the Whole Genome Shotgun (WGS) strategy and are therefore expected to be less complete than the six model organism genomes that were mostly sequenced in a hierarchical 'clone-by-clone' fashion. Despite variations in the completeness of these WGS genomes, the results of the mapping procedure were comparable to the results from the initial set of six genomes and orthologs of most of the 458 core

Table 3. Accuracy of CEGMA pipeline in recently sequenced genomes

Genome	Mapped (exons)	Genes cDNA	Nucleotide	Internal	Donor	Acceptor	Start	Stop	CDS
A.gambiae	453 (1539)	20 (69)	98.9/98.0	99.6/82.9	99.6/89.1	98.0/87.3	75.0	85.0	60.0
C.intestinalis	433 (3062)	25 (138)	97.4/98.9	94.4/97.7	93.8/96.4	95.6/98.2	72.0	98.5	72.0
C.reinhardtii	407 (3551)	31 (244)	95.0/97.1	91.4/94.9	91.1/94.2	93.4/96.6	74.2	93.5	61.3
T.gondii	303 (1946)	45 (235)	96.4/97.3	93.3/92.7	96.8/96.3	93.7/93.2	84.4	84.4	66.7

The number of core genes mapped in each genome. The number of exons is showed in parentheses. Gene with cDNA corresponds to the number of mapped genes for which we found a complete cDNA. The accuracy measures are computed only in the subset of genes with cDNA. Values shown refer to sensitivity and specificity (Sn/Sp), all numbers are percentages.

genes were mapped (Table 3). The worst mapping results occurred for *T.gondii* where 303 core genes were found. This species is a parasitic protozoan in the phylum Apicomplexa that diverged from metazoan/fungi/plantae around 1600 Mya. Closer inspection of the *T.gondii* genome revealed that some potential orthologs were missed by our protocol due to the high degree of sequence divergence (data not shown).

To assess the accuracy of the genes that were mapped, we utilized a set of complete gene structures that are supported by experimental data. Many of these were collected by Lomsadze *et al*. (2005) but others were added from a more recent release of GenBank (release 156). To use these genes as a test set, we first identified which of these genes corresponded to the core genes that were mapped. The overlap between the two sets produces a small subset of genes (20 *A.gambiae*, 25 *C.intestinalis*, 31 *C.reinhardtii* and 45 *T.gondii*) with experimental data that were used to measure the accuracy of predicted orthologs. In all four species, the predicted orthologs have highly accurate gene structures with >95% of all coding nucleotides correctly predicted (Table 3). Overall, these results indicate that the mapping procedure should be applicable to a wide range of genomes. Even for a species with very highly divergent genes (*T.gondii*), we still reliably map 66% of the set of core genes and predict gene structures with an accuracy as high as for the other species.

3.4 Mapping core proteins into draft genomic sequences in different stages

To assess the utility of the mapping protocol in genomes in draft stages, we also analyzed the number of core genes that CEGMA was able to map in WGS genomes at different levels of sequence coverage. For this purpose we used the different versions of the *T.gondii* genome available from the ToxoDB database (http://www.toxodb.org). This allowed us to compare six different genome assemblies with levels of sequence coverage ranging from 0.7× to 10×. The number of mapped genes increases with the increasing coverage of each genome assembly. At 2× coverage we map a third of the final number of mapped genes rising to two-thirds in the 4× assembly (Fig. 4). At higher levels of sequence coverage (6× and 10×) there is no great difference in the number of genes that we map, which might suggest that most of the genome sequence is represented in the 6× assembly.

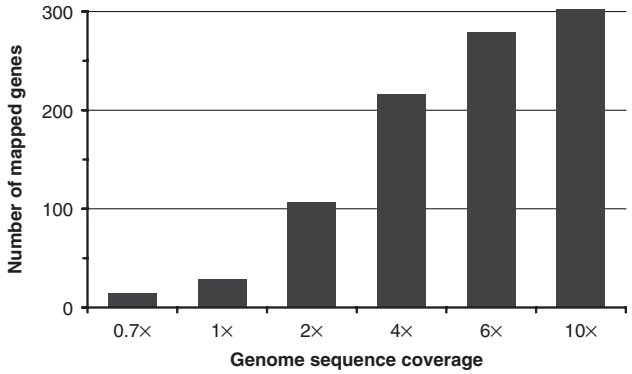

Fig. 4. Numbers of core eukaryotic genes mapped by CEGMA in *T.gondii* genome assemblies of different sequence coverage. Numbers of mapped genes are from a set of 458 core genes.

4 DISCUSSION

In this study, we define a set of 458 core proteins that are highly conserved in a wide range of eukaryotes and we present a procedure, CEGMA, which allows one to map the exon–intron structures of these core proteins to a new genomic sequence. The average accuracy achieved by CEGMA is 98% at nucleotide level and 90% at internal exon level in the six model organism genomes. CEGMA also produces similar accuracy in the four recently sequenced genomes (for the subset of predictions supported by complete cDNAs). An important advantage of our method is that we do not need to have any knowledge of the target genome other than the genome sequence itself. Our goal is not to annotate an entire genome using this protocol but instead to generate a highly reliable set of genes for the initial steps of annotation in new genomes. Our method is fast (1 day for the human genome using a Macintosh Quad-core G5 2.5 GHz) and accurately predicts hundreds of gene structures in any eukaryotic genome. We show that it can additionally be used in draft genomes, and even in a very diverged genome (*T.gondii*) CEGMA mapped 234 core proteins in a low coverage (4×) genome assembly. This is important because there is an increasing number of genome projects which are only being sequenced at low levels of sequence coverage (such as a range of 2× genomes in the Ensembl database).

Having shown that we can reliably find the orthologs of core genes in a new genome, we can consider whether these genes are suitable to train gene finders or to derive parameters for semi-automatic annotation pipelines. We find that within each genome, the compositions of the start, stop and splice sites are nearly identical between core genes and an equal-sized set of randomly selected genes (Supplement #3). While core genes tend to have slightly shorter exons and introns, these features exhibit a high degree of variability (Supplement #4). It might be expected that highly conserved genes are likely to be highly expressed, and therefore to have biased codon usage (Akashi and Eyre-Walker, 1998) and this does appear to be the case (Supplement #5). However, any method of gene training that is based on ESTs or cDNAs would also be biased towards genes that are highly expressed. We are currently working on methods to train gene finders from biased training sets (manuscript in preparation).

There are other methods for annotating genomes that have no experimental data including the bootstrapping method (Korf, 2004) and a self-training method (Lomsadze *et al*., 2005). In the bootstrapping method, the initial training set is determined by a gene finder trained on one or several other 'well known' genomes. The self-training method is based on GeneMark (Borodovsky and McIninch, 1993). In this method, the unsupervised training progresses through several iterations till a point of convergence is reached where all the predictions are the same as that in the previous iteration. The performance of unsupervised models can be influenced by the presence of transposable elements that frequently carry genes required for their mobility. Furthermore, the accuracy of this method can vary substantially from genome to genome. Our method has the advantage that we always know what the final set of gene predictions should look like. Another advantage of core genes is that their structures are very likely to be correct. Therefore, in addition to their use as a training set, they can also be employed as a test set. The unsupervised methods described earlier have the problem that after the training iterations have finished, there is no way to assess how accurate the final predictions are. Our method would allow the core genes to be used to measure the accuracy of these automatic annotation methods.

The CEGMA pipeline can also be used to find core genes that are missing in the initial annotations of a new genome assembly. For instance, for the four recently sequenced genomes used in this study we have found some core genes that are not present in the existing annotations. For the genomes of *A.gambiae*, *C.reinhardtii* and *T.gondii*, we found no more than 10 core genes that were missing. However, in the Ensembl annotations of *C.intestinalis* we find 67 core genes that are not present (Supplement #6). This is surprising given the highly conserved nature of these genes. Whilst Ensembl does show cDNA sequences aligned to the correct regions for most of these missing genes, it's inability to predict the genes themselves is indicative of the limitations in many annotation pipelines in use today. A final advantage of our approach is that in addition to their use as gene prediction training and test sets, core genes can be used to gauge the completeness of a genome assembly (manuscript in preparation). A rough estimate can be computed as the fraction of the total set of 458 core genes that can be mapped.

The starting point for this work was the conserved groups of proteins in the KOGs database. This database is being expanded to include proteins from eight more eukaryotes (http://www.ncbi.nlm.nih.gov/COG) and this would enable us to build a more accurate set of core genes and even to build sets of core genes for specific phyla. The utility of the system for both of these purposes should increase progressively with the inclusion of new genomes, particularly those of early-branching eukaryotes. Using profiles for specific branches on the evolutionary tree could improve the accuracy of the method.

ACKNOWLEDGEMENTS

This research was supported by a grant from the National Human Genome Research Institute (HG K22-0064) to I.K. Funding to pay the Open Access charges was provided by NHGRI (National Human Genome Research Institute).

Conflict of Interest: none declared.

REFERENCES

Akashi,H. and Eyre-Walker,A. (1998) Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.*, **8**, 688–693.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Birney,E. *et al.* (2004) Genewise and genomewise. *Genome. Res.*, **14**, 988–995.

Borodovsky,M. and McIninch,J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems*, **30**, 161–171.

Brent,M.R. (2005) Genome annotation past, present and future: how to define an ORF at each locus. *Genome Res.*, **15**, 1777–1786.

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

Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Curwen,V. *et al.* (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.

Dehal,P. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Gelfand,M.S. *et al.* (1996) Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.

Grossman,A.R. *et al.* (2003) *Chlamydomonas reinhardtii* at the crossroads of genomics. *Eukaryot. Cell*, **2**, 1137–1150.

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

Holt,R.A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.

Kissinger,J.C *et al.* (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.*, **31**, 234–236.

Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Korf,I. *et al.* (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17** (Suppl. 1), S140–S148.

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

Natale,D.A. *et al.* (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica*, **108**, 9–17.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Parra,G. *et al.* (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.

Parra,G. *et al.* (2000) GeneID in Drosophila. *Genome Res.*, **10**, 511–515.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Yeh,R.F. *et al.* (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.