BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 757–763
doi:10.1093/bioinformatics/btr010

Genome analysis

Advance Access publication January 6, 2011

A novel hybrid gene prediction method employing protein multiple sequence alignments

Oliver Keller[1,*], Martin Kollmar[2], Mario Stanke[3,*] and Stephan Waack[1]

[1]Institute of Computer Science, University of Göttingen, Goldschmidtstrasse 7, [2]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen and [3]Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Strasse 47, 17487 Greifswald, Germany

Associate Editor: Alex Bateman

ABSTRACT

Motivation: As improved DNA sequencing techniques have increased enormously the speed of producing new eukaryotic genome assemblies, the further development of automated gene prediction methods continues to be essential.

While the classification of proteins into families is a task heavily relying on correct gene predictions, it can at the same time provide a source of additional information for the prediction, complementary to those presently used.

Results: We extended the gene prediction software AUGUSTUS by a method that employs block profiles generated from multiple sequence alignments as a protein signature to improve the accuracy of the prediction. Equipped with profiles modelling human dynein heavy chain (DHC) proteins and other families, AUGUSTUS was run on the genomic sequences known to contain members of these families. Compared with AUGUSTUS' *ab initio* version, the rate of genes predicted with high accuracy showed a dramatic increase.

Availability: The AUGUSTUS project web page is located at http://augustus.gobics.de, with the executable program as well as the source code available for download.

Contact: keller@cs.uni-goettingen.de; mario.stanke@uni-greifswald.de

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

Received on October 22, 2010; revised on January 1, 2011; accepted on January 4, 2011

1 INTRODUCTION

With ever faster and cheaper sequencing techniques, the amount of available nucleotide sequence data is growing rapidly. In comparison, the process of accurately annotating the generated data is still lagging far behind. Fully automated annotation is essential here as the sheer amount of data make manual inspection, even as a secondary step, impossible on the whole. Thus, improving gene prediction tools that perform automated annotation of protein-coding regions as accurate as possible is becoming increasingly important for the generation, for example, of the corresponding protein sequence data.

In eukaryotes, the computational identification of the *gene structure* (in simplified terms, the locations of the protein-coding exons in a nucleotide sequence) is a complex and error-prone task. The most direct approach is pursued by *ab initio* methods that do not need any input but the genomic target sequence itself. Commonly, sequences are considered the result of a random process, and the outcome is the gene structure that is most likely to randomly produce the target sequence. Parameters for the underlying probabilistic model, often a Generalized Hidden Markov Model (GHMM), are derived from a training set of gene structures verified as accurate. As in any probabilistic approach, the prediction accuracy is limited already by constraints inherent to the model, even if it is a perfect description of the data.

In order to overcome these theoretical bounds, it is necessary to employ extrinsic sources of information that give hints whether an interval of the sequence is, for example, a coding exon. These gene finding methods are usually based on alignments with informant sequences: *comparative* methods make use of similarities with the DNA of closely related species, *transcript-based* methods map sequences to the target genome that are known to be expressed (such as ESTs or RNA-Seq) while *homology-based* methods try to map transcripts of related genes. Current approaches combine these methods, some including an *ab initio* gene prediction. Gene structures found by these methods are the starting point for the pipelines generating reference annotations for genomes (Harrow *et al.*, 2009).

RNA-Seq (transcriptome sequencing with next-generation sequencing methods, Metzker, 2010) promises major advances for gene finding; however, currently the accuracy of RNA-Seq-based gene prediction suffers from mapping ambiguities and partially contained introns, especially in complex genomes.

Homology-based gene predictors, such as Genewise (Birney *et al.*, 2004) and Exonerate (Slater and Birney, 2005), can determine gene structures by mapping a single protein sequence to the target genome, others like Projector (Meyer and Durbin, 2004) combine homology-based and comparative approaches. Cui *et al.* (2007) presented a combined homology-based and comparative gene finding method that extends the prediction beyond the homologous part to a complete gene structure but requires an established gene structure for the informant sequence. Protein queries highly identical to the target sequence can be mapped with BLAT (Kent, 2002); the software Scipio (Keller *et al.*, 2008) can refine a mapping provided

*To whom correspondence should be addressed.

by BLAT such that the precise exon–intron structure of a gene is automatically recovered from the query.

The gene prediction program AUGUSTUS (Stanke and Waack, 2003) is able to incorporate *hints* from external sources to combine them with an *ab initio* prediction (Stanke *et al.*, 2006a, b, 2008). The method employed in AUGUSTUS is completely generic as to the source of the hints and the way they have been generated; in practice, however, they are almost always derived from alignment-based methods (Stanke *et al.*, 2006a, 2008), including the use of peptides from proteomics experiments (Castellana *et al.*, 2008).

At a later stage of the genome annotation pipeline, an important task is the classification of proteins into families and subfamilies based on sequence similarities. A correct classification obviously relies on accurate gene predictions. Conversely, membership to a family is a potential source of information that can be made available already to the gene prediction. This information is commonly stored in protein family databases that are easily accessible, for example via InterPro (http://www.ebi.ac.uk/interpro/, Hunter *et al.*, 2009), quickly growing and often equipped with precomputed models (also called *signatures*) in the form of profile-HMMs, multiple sequence alignments (MSAs) and similar representations. Furthermore, researchers specialized on specific families will be interested in tools that enable them to use their existing sequence repositories to improve the prediction on newly available nucleotide data.

While almost every resource of protein family signatures offers its own methods to classify *protein* query sequences supplied by the user (bundled by InterProScan, Quevillon *et al.*, 2005), these methods cannot be applied directly to (eukaryotic) *genomic* sequences, without the prior knowledge of the gene structure, and hence the coding sequence. On the other hand, most protein-based gene finding methods map single sequences rather than protein signatures. The program Genewise has an HMM mode that can perform a combination of gene prediction and protein signature recognition using a profile-HMM (of one family at a time) in place of a single protein sequence.

Here, we present a novel approach that uses what we call a hybrid method as it combines the existing *ab initio* model with the protein signature as an additional model for protein family membership of the resulting transcripts. It is implemented as an extension (the protein profile extension, PPX) to the program AUGUSTUS. In this approach, evidence from complementary sources, such as RNA-Seq, can be used simultaneously.

2 APPROACH

With the integration of a protein model into gene prediction, we pursue two goals: first, the identification of members of a given protein family and, more importantly, an increased accuracy of the prediction designed especially to improve identification rates.

In AUGUSTUS-PPX, protein families are modelled by *block profiles*. In this context, a *block* is an ungapped and highly conserved section of a MSA. The concept of a block was first introduced by Henikoff and Henikoff (1991), and then used to classify proteins in the Blocks database (Henikoff *et al.*, 1999; Pietrokovski *et al.*, 1996).

Very similarly, collections of blocks, together with the ranges of admissible distances between consecutive blocks, have been used as protein signatures, referred to as *fingerprint* (Attwood and Beck, 1994), and currently collected in the PRINTS database (Attwood *et al.*, 2003), a member database of InterPRO.

A block profile is a collection of position-specific frequency matrices, each describing the amino acid distribution in a block, similar to a profile-HMM. However, in contrast to profile-HMMs, sequence motifs modelled by blocks have a fixed length with no insertions or deletions permitted inside a block. Along the full length of a MSA, non-conserved regions alternate with blocks. These inter-block sequence parts are modelled in a block profile only by constraining their length.

In the extended version of AUGUSTUS, one block profile at a time can be provided as an additional input, representing a particular protein family of interest. Here, we will refer to it as a *protein profile* as it models a gene with respect to its protein sequence.

AUGUSTUS predicts genes using a GHMM, in which each of the states corresponds to the biological meaning of the sequence (exon, intron, intergenic, etc.); this turns a gene structure into a sequence of states, together with their sequence coordinates, also called a *parse*. The well-known Viterbi algorithm is used to compute the highest scoring among all possible parses. In a GHMM, the score corresponds to the joint probability that the model generates the target sequence using the parse.

The profile extension to AUGUSTUS evaluates, for each candidate gene structure, a similarity score of the predicted transcript to the profile, giving a bonus to genes matching the profile. While the gene prediction takes place in genomic space, the protein profile models the *protein* sequence that the predicted gene translates to.

Although profile-HMMs can be a more powerful sequence model, block profiles were chosen here as a protein signature because the integration into the existing GHMM requires a reduced complexity. This is compensated by the use of the *ab initio* model that can better predict the less conserved sequence parts.

The coordinates of the protein model are mapped to the input DNA sequence when considering a candidate gene structure. Genes consisting of multiple exons will induce a mapping of partial profiles, where blocks may be disconnected by introns. Inter-block regions impose a constraint on exon length between blocks, but modelling nucleotide composition in them is left to AUGUSTUS' exon model.

Genes that show no evidence for similarity to all of the blocks in the profile are predicted the same way they would without the profile extension. This is an advantage over purely homology-based approaches, which cannot predict regions without sufficient homology. On the other hand, exons containing distant blocks can be forced to belong to the same gene, addressing the split gene problem common to *ab initio* approaches: while the predicted coding regions of a long gene may largely agree, they are frequently mispredicted as several shorter ones.

Instead of using the output of a separate program as source of extrinsic information, as is the case in the hints approach introduced by Stanke *et al.* (2006b), the mapping of the block profile to the target sequence is created *in parallel* to the *ab initio* prediction, with a mutual interaction between both.

The profile can be complemented with information about conserved intron positions (relative to the protein sequence), among the members of a protein family. This *intron profile* is a type of

information that is already lost when dealing with MSAs but can be very valuable for the gene prediction.

3 METHODS

3.1 Block profiles representing protein families

3.1.1 Scoring function defined by block profiles A set of n blocks can be transformed into a set of frequency matrices, one for each block, containing the column-specific frequencies of amino acids.

The order of the blocks is assumed to be preserved throughout all sequences in the family. For each b, the interval $I_b=[d_b^{\min},d_b^{\max}]$ specifies the range of admissible distances between consecutive block motifs (or, in the cases $b=0$, $b=n$, between the first/last block motif and the sequence start/end, respectively).

We call such a collection of frequency matrices, together with the range intervals I_b, a *block profile*, in analogy to other profiles generated from MSAs. For the sake of brevity, we will use the term *block* also for the particular matrix that represents it. A block profile does not contain probabilities for insertions or deletions, and it does not model the sequence regions between blocks.

From each frequency matrix, *odd ratios* are obtained by dividing each entry by the random (background) distribution of amino acids, and used as the scoring matrix $R^{(b)}$ for block b. If $s=s_0\ldots s_{w-1}$ is an amino acid sequence, its similarity to block b of length w is expressed by the odds ratio score of s, the product of the respective entries of the scoring matrix: $\rho^{(b)}(s)=R_0(s_0)\cdot\ldots\cdot R_{w-1}(s_{w-1})$. More generally, a *partial block score* $\rho^{(b)}_{[j..k]}(s)=R_j(s_0)\cdot\ldots\cdot R_k(s_{k-j})$ can be defined if s is a (shorter) sequence of length $k-j+1$.

To classify a given protein sequence s of length w as a *block hit*, we turn the scoring function into a decision function by requiring $\rho(s)>\tau$, with a block-specific threshold $\tau=\tau^{(b)}$, which is controlled by global parameters θ_{sens} and θ_{spec} that give upper bounds for the expected error rates in the respective models for blocks and background sequences (for details see the Supplementary Material).

3.1.2 Generating block profiles To convert a MSA into a block profile that can be used as input to AUGUSTUS, we provide separate tools that compute frequency matrices from each conserved ungapped region in the alignment. By the definition of a block, each member sequence of the family must contain all motifs that are part of a block, without insertions or deletions. Only a minimum number (usually set to 6–10) of subsequent gapless columns will actually form a block. These columns we call *usable* for conversion to a block profile. Additional restrictions might be imposed, for example on the degree of conservation in a block column.

Large protein families may be composed of subfamilies characterized by domains that are shared only by a subset of all sequences. If no domain is present in all member sequences, a family cannot be described appropriately by one single block profile. A possible solution is to convert the alignment associated with each subfamily into a separate block profile. In the next subsection, we present a different approach of determining a 'core' alignment that can be transformed into a profile.

Once blocks have been extracted from an alignment, or retrieved from the PRINTS or Blocks databases, PSSMs can be calculated by determining column-wise relative frequencies. In our conversion scripts, we follow the methods used for the Blocks database: a position-specific weighting scheme (Henikoff *et al.*, 1990) is applied in order to avoid over-representation of similar motifs, and pseudo counts are determined by regularizing the position-specific counts with BLOSUM matrices. Range intervals are determined from an alignment by taking minimum and maximum lengths of inter-block sequence parts among all aligned sequences or taken directly from a database.

The restriction that no insertions or deletions occur in a block motif can be relaxed, either by splitting a block into two to allow insertions in a central position of a large block or by merging amino acid distributions of neighbouring positions. This way, also profile-HMMs (e.g. given in HMMER format)—also containing frequency tables, but equipped with deletion and insertion states—can be used for conversion into block profiles.

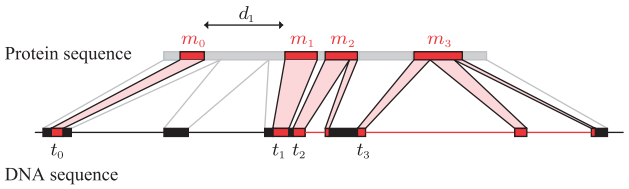

Fig. 1. A profile-DNA mapping. The translated transcript (above) contains four block motifs (in red) $m_0,\ldots,m_3$, separated by inter-block sequence (in grey) of length d_b. The coding gene (below) consists of six exons; sequences coding for the block motifs are shown in red. Block hits may appear inside an exon (m_0,m_1) or be disconnected by one or more introns (m_2,m_3).

3.1.3 Preprocessing of MSAs For the preparation of MSAs found in PFAM, we implemented an algorithm that iteratively discards sequences from an alignment until in both a required number of usable columns is reached (see above), and the estimated overall profile size (the number of usable columns multiplied with the number of used sequences) reaches a maximum. To this end, each column in an input alignment is categorized: Either

(1) it contains only a small number of gaps (removing the corresponding sequences would turn it into a block column) or

(2) it contains only a small number of non-gap characters (making it a candidate for complete removal from the alignment).

Columns that do not fall into these two categories are already determined to be inter-block columns at this stage, as well as isolated columns that cannot be extended to a block of minimal length.

Now, each column in the alignment can be considered as a potential start of a block of minimal length. Each sequence is *in conflict* with the block starting there if there is a deletion (the sequence has a gap in a column of Category 1) or an insertion (the sequence has a residue character in a column of Category 2) within the minimum number of columns after the fixed block start.

In each iteration, we use a heuristic to pick a new block start to be introduced by removing its conflicting sequence set from the alignment, based on the expected new profile size of the alignment; this is repeated until that size cannot be increased further by removing more sequences. The resulting subalignment can now be described by a block profile.

3.2 Evaluating profile-DNA mappings

A block profile is mapped to the DNA sequence σ by specifying the start locations $(t_0,\ldots,t_{n-1})$ of the segments coding for potential block motifs. We denote any such mapping by ψ. Since blocks may be interrupted by introns, the full mapping is well defined only when a candidate gene structure ϕ is also given, as depicted in Figure 1. In this case, ψ is *valid* if all t_b lie on exons belonging to the same gene, and the inter-block regions on the protein sequence have admissible lengths $d_b^{\min}\le d_b\le d_b^{\max}$.

The score of the mapping is defined by the odds ratio of the candidate block motifs m_b on the protein sequence

$$\rho(\sigma;\phi;\psi)=\rho^{(0)}(m_0)\cdot\ldots\cdot\rho^{(n-1)}(m_{n-1}). \quad (1)$$

Several genes may be equipped with mappings $\psi_1,\psi_2,\ldots,\psi_r$, resulting in a total score of

$$\rho(\sigma;\phi;\psi_1,\ldots,\psi_r)=\rho(\sigma;\phi;\psi_1)\cdot\ldots\cdot\rho(\sigma;\phi;\psi_r).$$

Formally, ρ can be defined to be 0 for invalid mappings. When provided with a protein profile, the Viterbi algorithm will search for the best-scoring

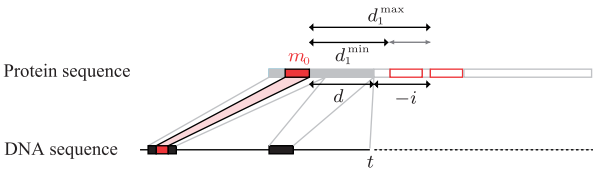

Fig. 2. Calculating Viterbi variables at DNA position t, for substate (b,i), $b=1$, $i<0$. The score is maximized for all gene structures that have a mapping to block motif m_0 at a fixed distance to the current position. In any parse continuing from this substate, the transcript must have a block motif m_1 starting between $d_1^{\min}$ and $d_1^{\max}$ from m_0. This determines the admissible range of block starts on the following exon (unless it is short enough to fit into the inter-block region). The bonus awarded to the Viterbi variable for this mapping is $\rho^{(0)}(m_0)$.

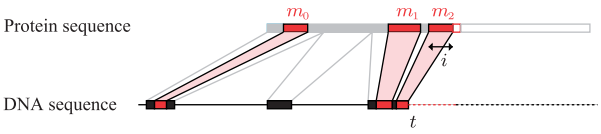

Fig. 3. Calculating Viterbi variables at DNA position t, for substate (b,i), $b=2$, $i>0$. The score is maximized for all gene structures that have a mapping to the truncated block motif m_2 of length i at the end of the exon. Any parse continuing from here must have the next exon starting with the remainder of the motif. The bonus awarded to the Viterbi variable for this mapping is $\rho^{(0)}(m_0)\cdot\rho^{(1)}(m_1)\cdot\rho^{(2)}_{[0..i-1]}(m_2)$.

combination of gene structure and compatible profile-DNA mappings. More precisely, it determines $\phi,\psi_1,\ldots,\psi_r$ maximizing the combined score

$$P(\phi,\sigma)\cdot\rho(\sigma;\phi;\psi_1,\ldots,\psi_r),$$

where the joint probability $P(\phi,\sigma)$ for gene structure and sequence is multiplied with the *profile bonus* $\rho(\sigma,\phi,\psi)$ for each candidate transcript equipped with a valid mapping ψ. Thus, only if a gene is compatible to a profile mapping with a score high enough to compensate for a lower *ab initio* score, the prediction with the profile will differ from the prediction without. In particular, on sequences where no members of the protein family are identified, the result will be identical to the *ab initio* prediction. For performance reasons, we consider for the evaluation of the profile bonus only mappings that consist of block hits (each of the $\rho^{(b)}$ must exceed their threshold $\tau^{(b)}$).

In the underlying sequence model, multiplying with ρ effectively amounts to replacing the background model with the block model, for the emission probabilities.

3.3 Integration into AUGUSTUS' state model

Iterating over all DNA positions, the Viterbi algorithm computes the scores of optimal partial parses (candidate gene structures) ending in the current position, and stores them in variables indexed by position and last state, each state representing a different sequence type, strand or reading frame. The state model AUGUSTUS uses has been described in Stanke and Waack (2003).

When equipped with a protein profile, the score assigned to a parse is modified as described above: for each gene in the parse that is compatible with a profile-DNA mapping, the score is to be multiplied with the best possible profile bonus; each exon in the gene contributes the bonus for the block (part) hits that overlap with it.

When the algorithm arrives at a position in the DNA, it must consider all partial profile mappings that started before and can be continued beyond this position, having a particular position within the profile aligned to the current DNA position (see Fig. 2). This position determines the conditions for the mapping to be continued; thus, a separate score has to be stored at that position for each profile position. To this end, the original state space of AUGUSTUS was extended by a set of substates attached to the main state, representing the position in the profile, specified as a pair of integers (b,i) denoting block (b) and position i relative to the block ($i\leq 0$ before a block, and $i>0$ inside a block). As shown in Figure 2, if i is negative, it determines the admissible range for the start of block b on the next exon.

The substates serve as additional indices for the Viterbi table labelling the entries that the *combined* scores are stored. For a partial profile mapping ending at the substate (b,i), the profile score multiplied to the *ab initio* score is given by Equation (1), except that the product is truncated at the current position: $\rho^{(0)}(m_0)\cdot\ldots\cdot\rho^{(b-1)}(m_{b-1})\cdot\rho^{(b)}_{[0..i-1]}(m_b)$ (the last factor is a partial block score included only in the case $i>0$, see Fig. 3). The maximum of all combined scores ending in the same substate is then the value stored in the table.

In general, substates can be used to model side conditions that influence the score of a parse, for example if constraints on the exon length are imposed, or for the exclusion of spliced stop codons. Here, substates are aligning the profile to the exon ends.

The emission probability of an exon in the extended version depends on the substates on both ends: it gets a bonus equal to the maximum score of all profile mappings on the exon that are compatible to the substates. Each pair of substates gives rise potentially to a different profile bonus. New substate variables are calculated by maximizing the combined score, over all predecessor positions and predecessor substates.

In introns, the emission probabilities remain unchanged, but intron states as well need to be equipped with substates to indicate a potential profile position they are mapped to.

Genes on the backward strand are evaluated essentially in the same way as genes on the forward strand; however, in this case, the profile is mapped to the DNA 'backwards', starting with the C-terminus. Correspondingly, the substates on the backward strand refer to blocks and columns in reverse order: the first motif is evaluated by the last block in the profile, starting with the last block column.

Since the model attaches a high number substates to every main state of the original model, an exhaustive evaluation of all combinations of block locations and substates would cause an explosion of running time and memory requirements; therefore, most of the substate entries are eliminated or shared between main states in order to control the computational cost. We store substate scores dynamically, reserving memory only for non-zero entries. See the Supplementary Material for details on speed-up strategies.

3.4 Fast preliminary block hit search

To make it possible a perform profile-based prediction on all relevant regions in a genome, a fast preliminary search algorithm was devised for determining the genomic regions that are likely to contain genes matching the profile, without determining detailed gene structures. The algorithm was implemented in a separate executable that can be used before running AUGUSTUS. Given a target sequence and a protein profile, it performs the following steps:

- Building a seed collection: for each of the 8000 possible triplets of amino acids, the positions in the profile are stored where that triplet is likely to occur.
- Determining block hit candidates: iterating over the target sequence, each 9-tuple is tested for being a seed. Any segment covered at least 25% by seeds for a block is considered for exhaustive evaluation. This prefiltering step is very fast, since it consists of a simple lookup with amino acid triples as keys.
- Exhaustive evaluation of candidates: given a block start offset, partial scores $\rho_{[j..k]}(a_j..a_k)$ are maximized where $a_0..a_{w-1}$ is the translated DNA segment. This interval is stored as a partial block hit if its size

reaches the minimal block width and the (partial) block threshold is exceeded.

- Assembling block hits to profile hits: each block hit is extended to a series of block hits by joining neighbouring hits, allowing for blocks being skipped. Using dynamic programming, sequences of block hits are determined that are highest scoring for all contained blocks, and returned as profile hits.

This algorithm is fast enough to be run on whole genomes with a high number of profiles. In order to determine the regions, AUGUSTUS-PPX is then run on.

4 DATASETS

4.1 Dynein heavy chain family

Dynein heavy chain (DHC) proteins belong to the longest proteins in eukaryotes comprising more than 4000 amino acids. They can be grouped into several subfamilies that are either responsible for intracellular transport and mitosis or part of the complex microtubule-based structures in cilia and axonemes. In mammals, most of the DHC genes are spread in up to hundred exons over several hundred thousand of base pairs. The 16 human DHC sequence fall into 10 subfamilies DHC1 to DHC9 and DHC11.

The DHC genes have all been manually assembled and verified because existing gene prediction programs were not able to correctly predict the gene structures over such long distances. For 13 of the 16 human DHC sequences, cDNA data are available and has been used for prediction. The remaining three DHC sequences, and the DHC genes of the other organisms, have been manually assembled based on comparative analysis of the metazoan homologs in the subfamily.

The manually created and curated MSA of the DHC family proteins, currently comprising more than 1600 DHC sequences, has been the best control and guide during this process. These manually assembled and verified DHC sequences are regarded as almost correct predictions and taken as reference sequences for the test runs. The DHC data are available from CyMoBase (http://www.cymobase.org, Odronitz and Kollmar, 2006).

From a complete sequence alignment of all DHC members available from CyMoBase, the alignment of the human sequences was converted into block profiles that were used as input in the test runs of AUGUSTUS.

A total of 96 DHC proteins out of 6 species were chosen as reference sequences; these were human, mouse, chicken, the Clawed frog and the Owl limpet, a sea snail. The range of members and subfamilies varies slightly between species.

4.2 PFAM alignments

To assess the performance of the profile extension on a wider set of protein families, five additional profiles were produced from MSAs downloaded from the PFAM database. The families were chosen randomly from the set of all families that had a minimum average length of at least 400 residues, and a minimum of 30 human representatives in them. Short alignments that cover only single domains were not considered, since AUGUSTUS-PPX is designed for the case of full-length protein signatures.

From each alignment, a core alignment was produced with the procedure described in Section 3.1. Typically, about 60% of the sequences were discarded in order to maximize the size of the usable part of the alignment. Among the core sequences, a total of 46 human sequences remained that were taken as reference and were downloaded from UniProt.

4.3 Genomic reference sequences

From the protein sequences chosen as reference set, i.e. the 96 DHC proteins from 6 species, and the 46 human sequences from the PFAM alignments tested, we produced reference gene structures. We used the program Scipio (Keller *et al.*, 2008; Odronitz *et al.*, 2008) to reproduce the exact exon/intron structure of the given proteins.

The reference genes for the DHC family consisted of 75 exons on average, spanning a length from 30 up to 450 kb. Depending on the quality of the assembly, frame shifts occur sporadically, and reference gene structures in some cases do not cover the full protein sequence. However, they can be considered complete with respect to the reference genome; the corresponding parts may have been determined from unmapped contigs or from cDNA analysis.

5 RESULTS AND DISCUSSION

5.1 Setup of runs

AUGUSTUS was run on the regions containing the reference genes, both in the *ab initio* and PPX versions, and the results were compared with the reference genes. The DHC genes were chosen as examples as their size makes their prediction prone to difficulties like the split gene problem. The human DHC used in the runs consisted of 42 blocks with a total length of 1214 sites (columns), the largest block having a motif length of 134.

In a second set of runs, allowing for the scenario that no ortholog to the target sequence is known, the human ortholog to the test sequence was removed from the alignment used to generate the profile, so that the remaining sequences all had <60% identity to the target sequence ('ex-ortholog').

We also ran Genewise in HMM mode on the reference regions, with default parameters and supplied with a profile-HMM in HMMER format, automatically generated from the same alignments, including the ex-ortholog ones. While Genewise has been succeeded by faster tools such as Exonerate in the case of single-protein queries, we are not aware of any other program that can perform a spliced alignment of a profile-HMM to a genomic sequence.

In order to compare a single-query mapping approach designed for high homology, cross-species searches with human queries for their orthologs were executed with Scipio, using default parameters.

Finally, to examine a random set of various reference proteins, runs were performed with the profiles generated from the five PFAM alignments, comparing AUGUSTUS-PPX to AUGUSTUS *ab initio* and Genewise. Here, we ran a low-similarity scenario by taking out from the alignments all sequences with >60% identity.

Apart from the block hit thresholds, no further tunable parameters were introduced to AUGUSTUS. The parameters have not been adjusted for the runs.

5.2 Assessment of the prediction quality

5.2.1 Results of the DHC runs The task of recognizing a gene as a member of the DHC protein family was accomplished by AUGUSTUS-PPX in almost all cases. Four sequences were not identified as DHCs, three from less similar subfamilies and one case with an incomplete genomic reference sequence. Instead, these

Table 1. Accuracy of DHC runs

Species	AUGUSTUS-PPX		AUGUSTUS *ab initio*	Scipio cross-species	Genewise	
	Full	Ex-ortho			Full	Ex-ortho
Highly accurate genes (%)						
Human	62.5	62.5	31.3	N/A	93.8	6.3
Mouse	72.2	61.1	22.2	88.9	55.6	0.0
Chicken	58.3	50.0	8.3	8.3	16.7	0.0
Frog	44.4	38.9	0.0	5.6	11.1	0.0
Zebrafish	57.1	35.7	57.1	7.1	14.3	0.0
Snail	44.4	44.4	5.6	0.0	11.1	0.0
Total	56.3	49.0	11.5	36.5	34.4	1.0
Exon level sensitivity (%)						
Range	86–94	84–91	78–85	32–86	59–80	46–50
Average	89.6	87.5	81.3	63.5	70.9	49.2
Exon level specificity (%)						
Range	76–92	75–91	77–88	52–88	74–85	68–72
Average	85.0	82.9	80.3	76.5	79.3	69.6

Comparing the results of AUGUSTUS-PPX with AUGUSTUS *ab initio*, single-sequence cross-species search (Scipio) and Genewise in HMM mode. 'ex-ortho' refers to runs with orthologs of target genes removed from the MSA. Highly accurate genes are those predicted with at least 95% sensitivity and 85% specificity.

genes were predicted identically to the *ab initio* version. Scipio (run only where human orthologs were there) and Genewise failed to identify two more sequences, and did not make a prediction in these cases. In the ex-ortholog scenario, Genewise left 14 sequences unidentified.

Table 1 shows prediction accuracy of the testing scenarios, compared with the *ab initio* version, Scipio and Genewise. At exon level, there is significant gain in sensitivity compared with the *ab initio* version, with almost 40% of the previously missed or mispredicted exons corrected, throughout the tested species. Specificity was overall increased slightly, but showed a decrease in some of the more distant species. Genewise is somewhat weaker in predicting exact splice sites, resulting in lower accuracy than AUGUSTUS *ab initio*, even when using the full profile.

With genes consisting of more than 70 exons, the requirement that a gene is predicted entirely correctly has to be relaxed somewhat when assessing prediction quality at the gene level. We calculated the rate of *highly accurate* genes, meaning that the overlap of prediction and reference is at least 95% of the reference and 85% of the prediction. On the human sequences, Genewise obtained the best results, with all genes but one predicted with high accuracy. In contrast to Genewise, AUGUSTUS-PPX maintained a high prediction rate on the more distant species and in the 'ex-ortholog' scenario. The protein-based prediction tools benefit especially from joining predicted exons to a single gene. The single-query-based Scipio predictions deteriorate strongly with evolutionary distance.

5.2.2 Effects of the protein extension There are various ways the profile can improve the prediction, as illustrated in Figure 4. A gene with complementary block hits on them, previously mispredicted as two genes, is now joined to one. This is an important advantage that protein-based gene finders have in comparison to *ab initio* tools.

Overall, specificity was improved to a lesser extent; in some genes, we observed a deterioration of specificity at exon level when

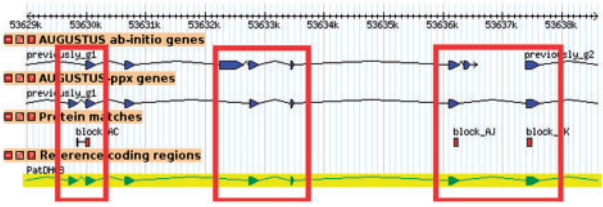

Fig. 4. A section of an example result of AUGUSTUS-PPX, shown in GBrowse, with the effects of the protein extension highlighted by a red box. One exon containing a block hit is added, one false positive exon removed to satisfy the distance constraints and two genes are joined into one.

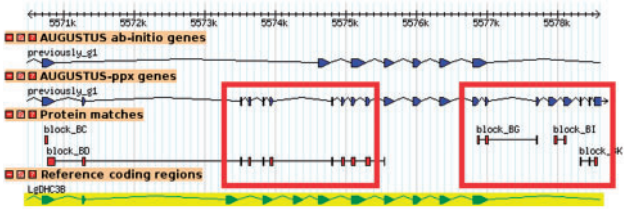

Fig. 5. Another GBrowse example, with false positive exons highlighted that are due to a missing block enforced by the profile extension.

using the profile, while in others it was clearly improved. These heterogenous results are caused by block hits enforced by the profile, leading to a number of false positive exons added to the prediction, as shown in Figure 5. This occurred both in cases of low homology and incomplete assemblies. We addressed this issue by performing a third set of runs based on the fast block search (see below).

5.2.3 Runs with profiles generated from PFAM alignments Results of the runs on the 46 reference genes from the five PFAM protein families are shown in Table 2. Exon level sensitivity and rate of highly accurate genes were improved, to varying extent, in all five cases. The number of completely correct genes rose from 10 (21.7%) to 14 (30.4%). In all but 1 of the 46 cases, the genes were identified by AUGUSTUS-PPX as members of their families. Results deteriorated only slightly when we restricted the profile to low identity, and occasionally even improved (removing similar sequences can lead to more blocks in the profile or prevent false positive block hits); two more sequences were not recognized as members. Genewise did not reach the accuracy of AUGUSTUS *ab initio*.

5.3 Further testing

5.3.1 Fast block search A fast block search (see Section 3.4) was performed prior to the actual AUGUSTUS runs, outputting the profile hits in the form of a list of blocks found there. Missing blocks were then removed from the DHC profile used for another AUGUSTUS-PPX prediction. With the filtered profile, no exons were added by enforced block hits, resulting in a higher exon specificity in all species (results shown as Supplementary Material). Generally, a fast block search is also needed to determine the regions for the gene prediction; it takes about the same time as the *ab initio* gene prediction, while the profile extension runs considerably slower (roughly 100 times for the large DHC profile). Run on the human

Table 2. Accuracy of PFAM runs

Family	AUGUSTUS-PPX		AUGUSTUS *ab initio*	Genewise
	Full	<60%		
Highly accurate genes at 95/85 (%)				
HSP70	87.5	87.5	75.0	0.0
Aldedh	77.8	55.6	66.7	0.0
AA_permease	53.3	53.3	33.3	0.0
Cullin	28.6	28.6	0.0	0.0
Sec1	28.6	42.9	14.3	0.0
Total	56.5	54.3	39.1	0.0
Completely correct genes (%)				
	30.4	28.3	21.7	0.0
Exon level sensitivity (%)				
Range	84–94	85–95	81–87	11–57
Average	89.0	88.7	88.8	39.3
Exon level specificity (%)				
Range	72–92	71–89	65–89	25–77
Average	80.5	78.9	75.9	57.7

AUGUSTUS-PPX was compared with AUGUSTUS *ab initio* and Genewise. '<60' refers to a profile with a maximum sequence identity of 60% to the target sequence, analogously to 'ex-ortho' above.

genome, no false positive hits were observed with the DHC profile; however, they cannot be excluded for shorter profiles.

5.3.2 Interoperability with external evidence We ran AUGUSTUS on the human sequences with manually edited hints and the DHC profile simultaneously. When supplied with the hints for the exons still missing in the original predictions, those were predicted correctly in general, unless there were non-standard splice sites. This showed that in principle there is the potential of adding the advantages of complementary methods, such as RNA-based evidence.

6 CONCLUSION

The gene prediction program AUGUSTUS was extended by a method combining protein-family based gene finding with an *ab initio* prediction. Equipped with protein signatures, prediction accuracy could be improved considerably, especially on full-gene level on very long genes. The extrinsic protein data significantly improves the gene prediction compared with existing programs when sequence data only from distant species was available.

The presented approach is complementary to transcript-based methods, and easily combined with them, offering the potential of a further improvement of prediction accuracy.

Block profiles are a protein signature suitable for aiding gene prediction. The approach for extending the model is generic and can be used to describe other types of constraints, for example on coding sequence length. Future plans include the integration of intron profiles, containing information about conserved intron positions.

ACKNOWLEDGEMENT

OK wishes to thank Shmuel Pietrokovski for sharing his knowledge about blocks and granting access to tools for evaluating them.

All authors thank Florian Odronitz for helpful discussions and comments.

Funding: Deutsche Forschungsgemeinschaft (KO 2251/3-1 and KO 2251/6-1 to M.K. and WA 766/6-1 to S.W.).

Conflict of Interest: none declared.

REFERENCES

Attwood,T.K. and Beck,M.E. (1994) Prints–a protein motif fingerprint database. *Protein Eng.*, **7**, 841–848.

Attwood,T.K. *et al*. (2003) Prints and its automatic supplement, preprints. *Nucleic Acids Res.*, **31**, 400–402.

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

Castellana,N.E *et al*. (2008) Discovery and revision of arabidopsis genes by proteogenomics. *Proc. Natl Acad. Sci. USA*, **105**, 21034–21038.

Cui,X. *et al*. (2007) Homology search for genes. *Bioinformatics*, **23**, i97–i103.

Harrow,J. *et al*. (2009) Identifying protein-coding genes in genomic sequences. *Genome Biol.*, **10**, 201.

Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19**, 6565–6572.

Henikoff,S. *et al*. (1990) Finding protein similarities with nucleotide sequence databases. *Methods Enzymol.*, **183**, 111–132.

Henikoff,S. *et al*. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.

Hunter,S. *et al*. (2009) Interpro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

Keller,O. *et al*. (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, **9**, 278.

Kent,W.J. (2002) Blat–the blast-like alignment tool. *Genome Res.*, **12**, 656–664.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

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

Odronitz,F. and Kollmar,M. (2006) Pfarao: a web application for protein family analysis customized for cytoskeletal and motor proteins (cymobase). *BMC Genomics*, **7**, 300.

Odronitz,F. *et al*. (2008) Webscipio: An online tool for the determination of gene structures using protein sequences. *BMC Genomics*, **9**, 422.

Pietrokovski,S. *et al*. (1996) The blocks database–a system for protein classification. *Nucleic Acids Res.*, **24**, 197–200.

Quevillon,E. *et al*. (2005) Interproscan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.

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

Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19** (Suppl. 2), 215–215.

Stanke,M. *et al*. (2006a) Augustus at egasp: using est, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.*, **7** (Suppl. 1), 1–8.

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

Stanke,M. *et al*. (2008) Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.