

Gene Organization Features in A/T-Rich Organisms

Karol Szafranski,¹ Rüdiger Lehmann,¹ Genis Parra,² Roderic Guigo,² Gernot Glöckner¹

¹ Institute for Molecular Biotechnology (IMB) Jena, Department of Genome Analysis, Beutenbergstr. 11, D-07745 Jena, Germany

² Grup de Recerca en Informàtica Biomèdica (GRIB), Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, Pg. Marítim de la Barceloneta 37-49, E-08003 Barcelona, Spain

Received: 19 December 2003 / Accepted: 18 August 2004 [Reviewing Editor: Dr. Manyuan Long]

Abstract. Several species have genomes in which the four nucleotides are not equally represented (Glöckner 2000). Interestingly, shifts to very high A/T or G/C levels can occur in several distinct branches of the tree of life. The underlying reasons for these shifts therefore may be of different origin. Now entire chromosome sequences from two different A/T-rich genomes, *Dictyostelium discoideum* and *Plasmodium falciparum*, are available (Bowman et al. 1999; Gardner et al. 2002; Glöckner et al. 2002). This gives us the opportunity to investigate how a high A/T content may influence the signals that are the landmarks for gene specification. We found that, in contrast with most known metazoan and plant genomes, splice signals contain, little information other than the canonical GT–AG dinucleotides. Intron lengths in A/T rich organisms, on the other hand, are comparable to those of other lower eukaryotes. Intergenic regions show, dependent on the orientation of adjacent genes, a size pattern with a ratio of 1 (3′–3′) to 2 (3′–5′) to 3 (5′–5′). Overall, gene organization patterns seem not to be influenced by the A/T bias. Surprisingly, the slightly higher A/T content of the *P. falciparum* genome compared to that of *D. discoideum* (80.1 versus 77.4%) is not achieved by increased A/T richness in intergenic regions. Instead both the shift of the nucleotide usage in coding regions to A/T-rich codons and the longer intergenic regions make an equal contribution to the higher A/T content in this organism.

Key words: A/T rich — Splice sites — Gene density — *Dictyostelium discoideum* — Genome organization

Introduction

To date several genomes in diverse evolutionary branches have been sequenced and many more will follow. Yet compared to their richness of diversity, lower eukaryotes are underrepresented in this selection. Thus, we are only at the dawn of the understanding of genomic organization in this large group of organisms. Here we attempt to describe common genomic features of lower eukaryotes with comparable high A/T contents. The different theories on why nucleotide biases in genomes occur were reviewed by Glöckner (2000).

The extraction of the genomic information content of the cell, in particular, the localization and structure of protein coding genes, is a challenge since no simple rules exist for this purpose. Even though the mechanisms by which gene structures are specified in eukaryotic genomes appear to be quite general, each genome seems to have its singular signature for gene recognition that was and is shaped by selective pressures during evolution (Gentles and Karlin 2001). This implies that gene structures and the transcription and translation apparatus in each species are coadapted to the same extent to enable the cell appropriate transcription and translation of each

gene. Deviations from a given signature may pinpoint sequence domains and genes that are introduced into a given genome from a foreign source (Karlin 2001). Thus, analysis of gene signatures of a species' genome not only provides tools for the most accurate prediction of gene structures but also may shed light on the common mechanisms underlying gene specification and its evolution. In addition, they may also be useful to resolve conflicting phylogenies and to track down horizontal gene transfers, genetic drift, and other evolutionary events.

Despite the sometimes obvious differences of foreign genes to the original genes of an organism, the cellular machinery is able to transcribe, translate, and use these genes seemingly without difficulties. Only in highly biased genomes artificially introduced genes with no adaptation to the host are not expressed (Fuhrmann et al. 1999). Computed probabilities for the existence and structure of a gene rely on the similarity of the gene structure to be detected to well-defined gene structures of a given organism. This statistical approach implies the prediction of false positives and inability to detect true genes. The ability to correctly use all the information content of a genome despite deviations of particular genes from the mean signatures distinguishes the cellular recognition of genes from gene detection by bioinformatic means.

Gene detection in prokaryote genomes is comparably easy. Here the open reading frames (ORFs) framed by appropriate translation start and stop signals very often equal the genes themselves. Yet overlapping genes and the potential of small ORFs to be meaningful for the cell are obstacles to a straightforward strategy for gene detection in prokaryotes. Thus, even in the simplest systems not all true genes can be detected and not all predicted genes are biologically meaningful. Therefore, gene prediction can only lead to an approximation of gene content and genome complexity.

Eukaryote genomes pose even more problems for gene detection. Here a gene is not only one block of information from translation start to stop but may be distributed over several exons separated by introns that can span several kilobases of genomic sequence. In addition, in contrast to the densely gene packed prokaryotic genomes, eukaryotic genes are separated from each other by large intergenic regions. These characteristics of eukaryote genes make it even more difficult to define correct gene structures than in prokaryotes. To overcome these limitations, eukaryotic gene prediction programs attempt to integrate the different species-specific gene signatures into likely candidate gene models (Hooper et al. 2000; Makarov 2002; Mathe et al. 2002; Rogic et al. 2001; Zhang 2002).

In the analysis presented here we focus on the analysis of *Dictyostelium discoideum* gene features

and gene predictions. This organism belongs to the amoebozoa, a branch believed to have diverged early in evolution, perhaps before the separation of animals and plants (Baldauf and Doolittle 1997; Baldauf et al. 2000). This species therefore may be best suited for the identification of ancient genomic features. Moreover, it has a compact genome that makes it easier to study gene features than in large genomes, where these characteristics may be hidden in large regions of, from our current point of view, seemingly senseless DNA. The genome examined is unusual with respect to nucleotide distribution since it is shifted toward a very high A/T content of nearly 80%. Why a genome is forced to develop such extreme nucleotide disparities is still unclear.

Extreme values of such biases in the same direction in different organisms may affect gene signatures in similar ways. To address this question we compare the results obtained with *D. discoideum* to those with *Plasmodium falciparum*, the causative agent of malaria. This organism has also a very A/T-rich genome (Gardner et al. 2002) that apparently evolved independently. This independent acquisition of the A/T bias is underlined by the deep branching of the organisms (Plasmodium, Alveolata > Apicomplexa > Haemosporida > Plasmodium; and Dictyostelium, Mycetozoa > Dictyosteliida > Dictyostelium) in the eukaryote phylogeny. Since we now have large amounts of A/T-rich sequences from these two highly biased genomes, we can identify common features of genes in these organisms. We here discuss also whether nucleotide biases may have a measurable impact on the evolution of gene features.

Materials and Methods

Gene Prediction Programs

The following gene prediction programs were used: geneid (<http://www.limim.es/geneid.html>), GlimmerM (http://www.tigr.org/tdb/glimmerm/glmr_form.html) trained for prediction in *P. falciparum*, and GenScan (<http://genes.mit.edu/GENSCAN.html>) trained for *A. thaliana* gene predictions. To measure accuracy we used the statistics described by Burset and Guigo (1996) and Reese et al. (2000).

Data Sets

To obtain a data set for the training of geneid, we extracted full-length entries of *Dictyostelium discoideum* genes from the GenBank database. The gene structures should be derived from nuclear genes, not being of mitochondrial origin. Furthermore, each gene we included in the data set was unique, i.e., it did not share more than 80% identity on the nucleotide level with other genes in the data set. The final *Dictyostelium discoideum* training set contains 140 genes, 38 corresponding to single-exon genes and 102 multi-exonic genes containing 180 introns. The genes, which were predicted on chromosome 2 after training of geneid with the above set, were then used for further analysis.

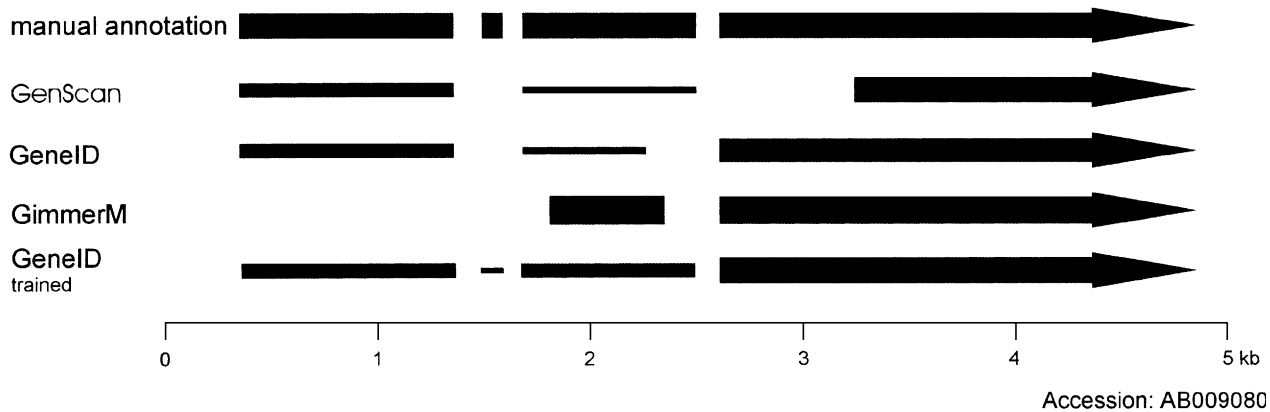


Fig. 1. Comparison of gene prediction programs. The programs were GeneScan, GlimmerM, untrained geneid, and trained geneid. The figure shows the analysis of the *D. discoideum* genomic sequence accession No. AB009080 representing the *D. discoideum* gene product TRFA. Where applicable the score values for predicted exons are given as the height of the rectangles representing the exons.

The *P. falciparum* data set was constructed from the annotated sequence of chromosome 3. This annotation was carried out by the Sanger Institute group using manual curation of the output of different software packages (Hall et al. 2002). The sequence corresponding to this chromosome contained 234 genes, 98 corresponding to single-exon genes, and 136 multiexonic genes containing 379 introns.

Splicing and Translational Signals

Donor and acceptor sites were extracted from the *D. discoideum* chromosome 2 data set. The information content of each position around the signal was computed as by Burge et al. (1999). The positions taken for each site were -2 to $+5$ for donors -15 to -1 for acceptors, and -6 to $+5$ for the translation start side (0 being the first position downstream between coding and noncoding junction and -1 being the first position upstream). Background probabilities have been computed in the same positions around every minimal conserved nucleotide for each site (corresponding to GT for donors, AG for acceptors, and ATG for start sites) in the training set sequences not annotated as functional sites. The final score for each site was computed as a log-likelihood ratio of the frequencies of the real annotated sites versus the frequencies of the bases around the nonannotated minimal sites (assuming that the nonannotated sites in these sequences correspond to the distribution of the nonreal sites). Compositional properties were computed as by Guigo (1999) and Parra et al. (2000).

Evaluation of available prediction programs for their ability to predict correct *D. discoideum* gene structures was performed using the training set. All data sets used in this study with geneid and the parameters inferred are available at <http://genome.imim.es/datasets/D.discoideum/> and <http://genome.imim.es/datasets/P.falciparum/>.

Results

Gene Prediction in *D. discoideum*

Investigation of determinants of gene structures requires an experimentally confirmed data set consisting of as many complete gene structures as possible. For *D. discoideum* we obtained a data set of 140 genes after filtering for a number of features (see Materials and Methods).

With this data set we examined the performance of three gene prediction programs without prior training geneid (Blanco et al. 2003), GenScan (Burge and Karlin 1997) with *Arabidopsis thaliana* parameters, and a version of GlimmerM (Salzberg et al. 1999) designed for gene prediction in *Plasmodium falciparum*. We observed that the three programs worked similarly well on this data set, with GlimmerM being more sensitive than geneid, and geneid more sensitive than GenScan (supplementary material).

Next, we tested a version of geneid specifically trained for gene prediction in *D. discoideum*. This version showed great accuracy, with 99% of the coding nucleotides correctly predicted and 97% of the nucleotides predicted as coding actually being coding (that is, at the nucleotide level, sensitivity was 0.99 and specificity 0.97). With the trained version of geneid, even small annotated exons missed by the other programs could be detected (Fig. 1). The accuracy of the geneid predictions in *D. discoideum* is clearly superior to that in human (here the sensitivity as measured in human chromosome 22 is 0.8 and the specificity 0.63; data not shown).

Predicted Genes

On chromosome 2 of *D. discoideum* geneid predicted 2799 genes (>40 amino acids [aa]) (Glöckner et al. 2002). Since the statistical probability to be a true gene increases with the length of the predicted gene, the low threshold chosen may have led to a slight overestimation of gene numbers. Of the predicted genes 109 (3.7%) were incomplete at the 5' end and 56 (1.9%) were incomplete at the 3' end. Since EST data for *D. discoideum* are scarce, especially for the 5' end, we could only determine the real gene structure for a minor part of the predicted incomplete genes. In many cases the lack of completeness at the 5' end may be due to very short portions of coding sequence on

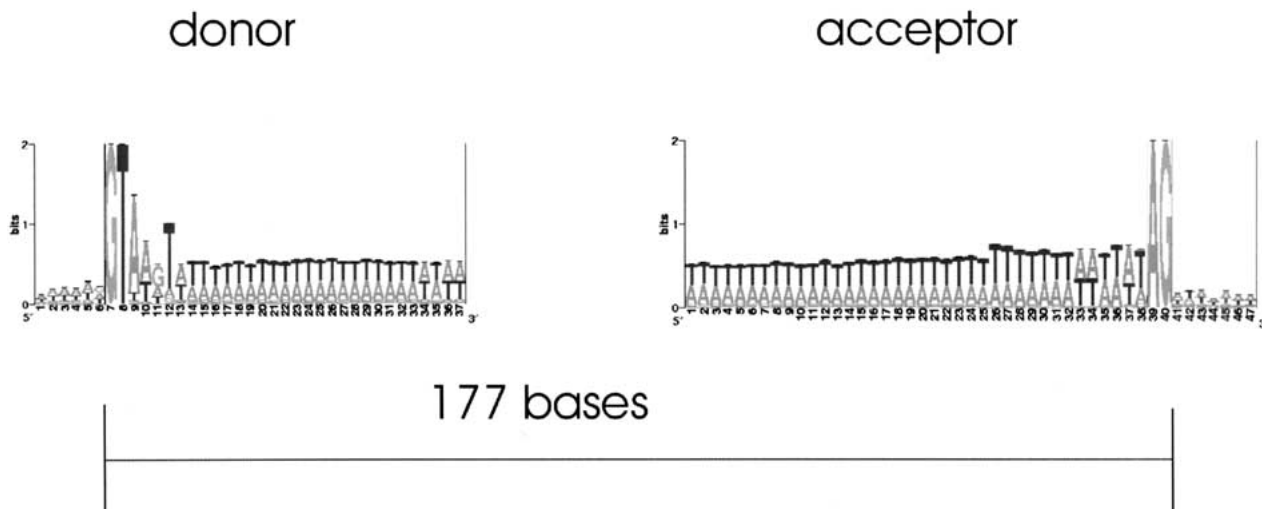


Fig. 2. Splice site signatures in predicted genes of chromosome 2 from *D. discoideum*. The height of the base letters at each position is proportional to their respective occurrence at this specific position. The black line between donor and acceptor site signature images gives the median intron length. Additional figures are given in the supplementary material.

the first exons. Incompleteness at the 3' end results from artificial prediction of a splice donor site shortly before the real stop codon. Generally, the analysis of the incomplete predicted genes showed that the bordering sequences exhibit a very high A/T content in which no obvious coding potential could be detected. Thus, we conclude that only very small portions of the coding region of these partially predicted genes are missed.

Exon–Intron Structure, Splice Signals, and Intergenic Regions

On average the introns of the predicted genes from chromosome 2 comprise 177 bases. *D. discoideum* genes also have a relatively small number of introns per gene, comparable to that of *Schizosaccharomyces pombe* but more than that of *Saccharomyces cerevisiae* (Deutsch and Long 1999; Kriventseva and Gelfand 1999; Wood et al. 2002). Interestingly, this is a significantly higher value than that of the training set, with 132 bases on average. The average number of introns per gene is 1.28. The maximum number of introns in one gene was 15; overall 48 genes (1.7 %) have more than 4 introns.

In contrast to most known metazoan and plant genomes, *D. discoideum* splice signals carry little information other than the GT–AG canonical dinucleotides (Fig. 2). Only the A nucleotide in position +2 of the donor site (+0 being the first position in the intron, occupied by the canonical G) and a G at position +4 appear to be clearly overrepresented with respect to the expected values.

Due to the high number of genes in a relatively small genome, the intergenic regions are short. Here we count the entire space between putative translated

Table 1. Comparison of intercoding regions (ICRs) between different species

Species	Orientation of adjacent genes (left–right)	Number of intergenic regions	Mean size	Median size
<i>D. discoideum</i>	3'–5'	1427	704	560
	3'–3'	752	331	246
	5'–5'	712	963	815
<i>P. falciparum</i>	3'–5'	96	1966	1540
	3'–3'	61	1079	810
	5'–5'	60	2830	2463
<i>S. cerevisiae</i>	3'–5'	357	564	409
	3'–3'	208	350	249
	5'–5'	186	642	500
<i>A. thaliana</i>	3'–5'	2170	2416	1340
	3'–3'	1004	2001	708
	5'–5'	1006	2981	1958

gene loci from the start or stop signals of one gene to that of the next as intergenic, not taking in account the untranslated but transcribed regions (UTRs) of a gene. On chromosome 2 these “intercoding” regions (ICRs) have an average length of 786 bases on chromosome 2. A similar average size was found for the ICRs of chromosome 1. We therefore think that these values are valid for the whole genome. We found that the average size of ICRs differs dependent on the orientation of the adjacent coding regions (Table 1). The translated portions of the genes with neighboring 5' ends are separated from each other by three times more space than translated portions of genes with neighbouring 3' ends. We then tried to estimate the length of the untranscribed intergenic regions using available EST data and a set of manually annotated genes. Of the 36 genes analyzed, 18 were randomly chosen, and 18 belong to the rasGEF family of proteins (supplementary information). We

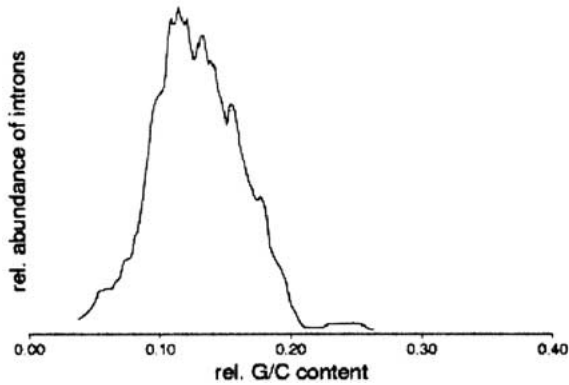
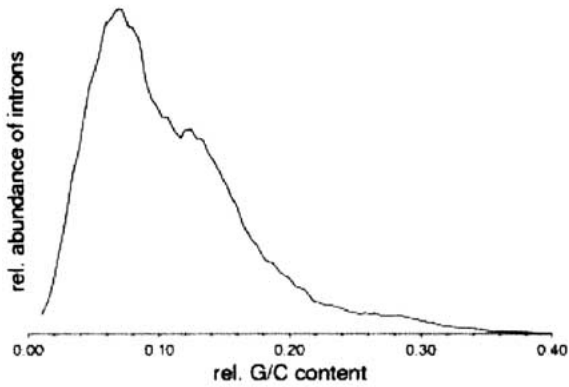
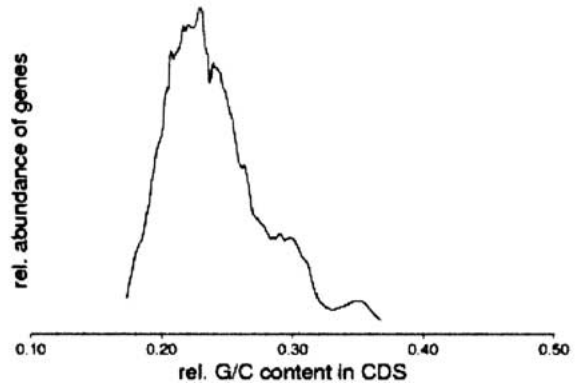
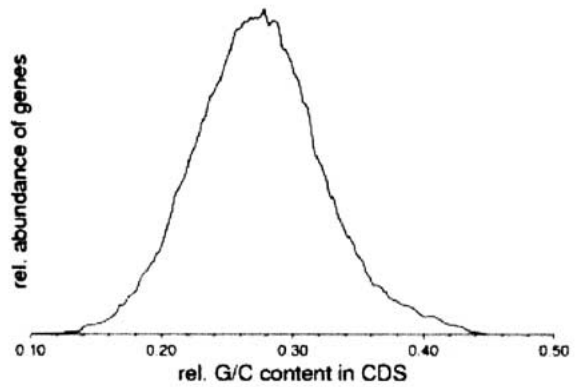
A) introns**B) CDS***D. discoideum**P. falciparum*

Fig. 3. G/C distribution in introns and coding regions of *D. discoideum* and *P. falciparum*. (A) Distribution in introns. (B) Distribution in coding regions.

found that the 36 genes tested have a 3'UTR of 74 bases on average from the end of the coding region to the poly(A) tail, with a standard deviation of 58 bases. Since ESTs of *D. discoideum* do not represent the 5' ends of the genes very well, we did not analyze the 5'UTR lengths.

Comparison to Organization in Other Species

To analyze the influence of nucleotide biases on gene properties we compared the results obtained with *D. discoideum* to a second A/T-rich genome, that of *Plasmodium falciparum*. For this purpose we used the previously annotated chromosome 3 (Bowman et al. 1999). We found the same restriction of the splice signal to the canonical GT-AG motifs with some weak preferences for certain nucleotides near the canonical motifs (supplementary material). This may be partially caused by the extreme A/T values. The mean intron length of 209 bases as well as the A/T content of 87% is in the same range as *D. discoideum*.

Figure 3 shows the G/C distribution in introns and coding regions in *D. discoideum* and *P. falciparum*. Interestingly, in coding regions of *D. discoideum* the G/C values follow a Gaussian distribution, whereas

the *P. falciparum* distribution is characterized by a long extension toward higher G/C values.

The gene density in *P. falciparum* is much lower than that of *D. discoideum*. Here the mean distance between the coding regions is around three times the distance in *S. pombe* and *D. discoideum*. Since the mean gene and intron lengths differ only slightly between both organisms, only different mean lengths of ICRs are causative for the different observed gene densities. We found that as in *D. discoideum* the average size of ICRs differs dependent on the orientation of the adjacent coding regions (Table 1).

Interestingly, for both organisms the A/T content of ICRs is the same, at 86%. In contrast, the coding sequences (CDS) in *P. falciparum* exhibit a more extreme bias toward the usage of codons with A and T nucleotides (data not shown). This preference results in a lower G/C content in CDS in *P. falciparum* (24%) than in *D. discoideum* (27%).

We examined the ICRs of two further organisms, *Arabidopsis thaliana* and *S. cerevisiae*. The plant species exhibits a gene density comparable to that of *P. falciparum*, at 4.4 kb/gene (The A. thaliana Genome Initiative, 2000), albeit *S. cerevisiae* constitutes an extreme, at 1.8 kb/gene (Goffeau et al. 1996). The

Table 2. Performance of GeneID in A/T-rich organisms

Species	Parameter file	SN	SP	CC	SNe	SPe	SNSP	ME	WE
P.f.	P.f.	0.98	0.98	0.95	0.70	0.71	0.70	0.13	0.11
D.d.	D.d.	0.99	0.97	0.97	0.76	0.75	0.76	0.06	0.06
P.f.	D.d.	0.71	0.96	0.68	0.21	0.18	0.19	0.19	0.42
D.d.	P.f.	0.96	0.96	0.86	0.52	0.52	0.52	0.17	0.16

Note. SN—sensitivity (nucleotide level); SP = specificity (nucleotide level); CC—correlation coefficient; SNe—sensitivity (exon level); SPe—specificity (exon level); SNSP = (SNe + SPe)/2; ME—missing exons (fraction of true exons not predicted); WE—wrong exons (fraction of predicted exons not overlapping any true exon); P.f.—*Plasmodium falciparum*; D.d.—*Dictyostelium discoideum*.

A/T content of both species does not reach the extremes as in *D. discoideum* or *P. falciparum*. For our analysis of ICRs we used *A. thaliana* chromosome 4 and *S. cerevisiae* chromosome 4. As can be seen Table 1 the ICR length depends on the gene orientation in all genomes irrespective of the gene density. The ratio of the ICR lengths can be roughly given as 1 (3'/3'):2 (3'/5'):3 (5'/5') for all organisms but *S. cerevisiae*, if we take the median instead of the mean values into account (Table 1).

Discussion

Since research on *Dictyostelium discoideum* focuses mainly on cytoskeleton, signal transduction, and development, the described genes and therefore the gene features of the training set for this organism may not represent the average gene of this organism. Previous studies are limited by the preselection of certain gene families (Rivero 2002). Future gene predictions and genome analyses have to include also genes from the primary metabolism and other functional groups. This way, gene prediction will become more accurate and efforts at manual curation of the predicted gene data set will be minimized. Despite these limitations the analysis presented here gives some insight into the gene organization of lower eukaryotes with a high A/T bias.

A bias toward a preferred use of A and T nucleotides can be found in several organisms belonging to well-separated branches in the tree of life (Glöckner 2000). These organisms evolved these more or less pronounced biases in the same direction independently. This genomic bias also shows local effects in that it overlays the gene features that are common to all organisms. The *D. discoideum* genome exhibits one of the most pronounced biases, with an A/T content of more than 78%. This value is only surpassed by that of *P. falciparum*, with an A/T nucleotide content of 80%.

Despite the clear similarities in gene specification and gene organization between the two A/T-rich genomes, gene signatures still appear to be quite specific. Table 2 shows the accuracy on *P. falciparum* sequences of the version of geneid trained on

D. discoideum sequences, and vice versa. Surprisingly, predictions using the *D. discoideum* version of geneid on *P. falciparum* sequences were substantially worse than predictions using the *P. falciparum* version of geneid on *D. discoideum* sequences. This could be attributed to the more Gaussian distribution of G/C content in *D. discoideum*, indicating more homogeneous gene features in this species than in *P. falciparum*. A/T richness alone thus may not be a key feature that influences gene structure signals.

Cells are also able to transcribe and translate foreign genes if they do not have a very different base composition as can be seen from cross-species complementation. Evolutionary constraints therefore led to a well-conserved transcription/splicing/translation machinery. This conservation should enable a general approach for gene detection in unknown sequences. Yet currently available software tools have to be individually trained for a given organism for optimal performance. This analysis shows that, even if extreme compositional biases are shared between organisms, training for the specific organism is indispensable. Thus, gene recognition in the cell is different from bioinformatic approaches and may use more widespread, not yet detected signals.

Like all other lower eukaryotes for which sequence data are known, *D. discoideum* has only short introns. Surprisingly, the mean intron length of the training set was 132 bases, much shorter than that of all chromosome 2 genes (177 bases). This difference is in part due to predicted long (>1000 bases) introns in cases where geneid was not able to resolve the correct gene structure. Manual inspection of all 74 introns larger than 1000 bases revealed that at least 90% of them are artifacts. These artificial introns are the main source of the observed difference between intron lengths of the training set genes and those of the chromosome 2 genes.

Although introns are small and rare in the *D. discoideum* genome, they could slightly contribute to the overall A/T bias of the genome. The canonical GT-AG motif at the donor and acceptor sites of introns is conserved with only small deviations throughout the tree of life. In *D. discoideum* the requirements for proper splicing seem to be almost entirely restricted to this common motif. Besides

these common sites, only weak preferences for A nucleotides downstream of the donor site could be detected in *D. discoideum*. The further positions are only slightly favored by a (A/T)GT motif. A similar picture was obtained when analyzing *P. falciparum* splice sites (supplementary material). This is in contrast to metazoa and plants, where additional nucleotides strongly contribute to a consensus motif. The splice apparatus therefore has to be able to correctly detect and process these signals in spite of the relative weakness of the signals compared to other organisms. Possibly the difference in A/T content between intron sequences and coding sequences as can be deduced from the different nucleotide compositions is used by the cell as additional signal. Due to their lack of special requirements concerning sequence motifs, the entire intron sequence is able to contribute to the high mean A/T content in the genome of 78%.

For *S. pombe* as well as for *S. cerevisiae* it was shown that there is a strong bias of intron positions toward the 5' end of the genes (Sakurai et al. 2002). This is also the case for the introns of *D. discoideum*. The mechanism by which the preferred 5' localization of introns is achieved is not yet entirely understood (Fink 1987; Mourier and Jeffares 2003; Wood et al. 2002). Considering the early branching of the mycetozoa, to which *D. discoideum* belongs (Baldauf and Doolittle 1997), this mechanism may be established relatively early in evolution.

ICRs in *D. discoideum* are relatively short. The length of these regions depends also on the orientation of the adjacent genes (Table 1). On average half of the space between tail-to-tail-oriented coding regions is occupied by the 3'UTRs of the genes, as could be identified using available EST data. Thus, the real mean distance between transcribed units in the tail-to-tail orientation is only 150 bases. The situation for 5' ends remains unclear since too few confirmed full-length mRNA sequences are available. But analyses from other organisms suggest that 5'UTRs are shorter than 3'UTRs (Jareborg et al. 1999). If this is also true for *D. discoideum* genes, the function of the additional space of 300 bases/gene required for genes in the head-to-head orientation cannot be attributed to long 5'UTRs.

The orientation-dependent characteristic of intergenic spacer regions was also described for *S. pombe* (Wood et al. 2002), an organism with the same high gene density as *D. discoideum*. The respective values for both organisms are in the same range, thus the contribution of gene orientation to the size of intergenic regions is independent of nucleotide biases. Interestingly, in *S. cerevisiae* with the highest gene density among eukaryotes measured so far, the size of the ICRs is practically not correlated with the orientation of adjacent genes. This may be due to the fact that this organism exhibits the highest gene

density so far measured and therefore a minimum of ICRs.

P. falciparum has a considerably lower gene density than *D. discoideum* or *S. pombe*, but the CDS and introns have a comparable size. Thus, longer ICRs, which occupy around three times as much space in *P. falciparum* as in *D. discoideum*, mainly cause the lower gene density. For the ICRs we also found a gene orientation-dependent size (Table 1). Interestingly, for both organisms the ratios for the orientations 3'-3', 3'-5', and 5'-5' follow the rule 1 to 2 to 3. The same holds true for the *A. thaliana* ICRs, if we calculate the median of the ICR sizes. Thus, for gene densities between 2.5 and 4.8 kb/gene there seems to exist a rule for proper gene spacing. The deviation of the mean values from this rule may reflect the more relaxed gene spacing in this genome. In fact, for the human genome with its very large ICRs, we could not detect such a rule (data not shown). *S. cerevisiae* as the other extreme shows compressed ICRs and also does not follow this rule.

If we view a genome as concatenated blocks of gene regions, we can conclude that the 5' region down to the start codon of such a block occupies two times as much space as the 3' region from the stop to the end, irrespective of the absolute length. Analysis of further compact genomes, preferably of diverse protists, is required to determine whether the 1:2:3 size distribution plays an important role in genome organization or is an intrinsic property of a gene block.

To achieve a higher A/T content than the *D. discoideum* genome *P. falciparum* would have had several possibilities: (i) extension of intergenic regions, (ii) elevated A/T values in distinct sections of the genome, and (iii) higher A/T values all over the genome. The 3% higher overall A/T content of *P. falciparum* compared to *D. discoideum* can be attributed to the first and second possibilities, since CDS show a higher A/T content, and the intergenic regions are three times as long as in *D. discoideum*, but show the same mean A/T content of 86%. Even if the ICRs of *D. discoideum* were adjusted to match the length of ICRs in *P. falciparum*, the A/T content of this genome would be only 78.8%. Thus, about half of the higher A/T bias in the *P. falciparum* genome can be attributed to the longer intergenic regions; the other half is caused by the nucleotide bias in CDS. It is not easy to understand why, in the case of *P. falciparum*, a contribution of CDS nucleotide bias to the overall genome bias was preferred over elongation or achievement of higher A/T values for intergenic regions. Due to the need to retain functionality, coding sequences commonly underlie stronger evolutionary pressures than intergenic regions. If natural selection were the cause for nucleotide biases in genomes, the organism would tend to minimize the costs of this

shift. Therefore, one could speculate that a compact genome was preferable to extended intergenic regions. Additionally, the value of 86% A/T might be the upper threshold which can be reached over several kilobases of DNA.

Conclusion

Gene recognition patterns seem to be conserved over large phylogenetic distances, but they also appear to have some taxon-specific component. Moreover, gene structure properties are only slightly influenced by nucleotide biases in a genome. Intron and intergenic size distributions show no differences in A/T-rich versus normal genomes. Furthermore, space requirements before and after genes seem to follow a strict rule which is only weakened in large or very compact genomes. Thus, A/T richness is only one additional feature of a genome that is superimposed on regular features found in all genomes.

Acknowledgment. We thank Matthias Platzer for carefully revising the manuscript.

References

- The Arabidopsis thaliana Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Baldauf SL, Doolittle WF (1997) Origin and evolution of the slime molds (Mycetozoa). *Proc Natl Acad Sci USA* 94:12007–12012
- Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290:972–977
- Blanco E, Parra G, Guigo R (2003) Using GeneID to identify genes. In: Baxevanis A (ed) *Current protocols in bioinformatics*. John Wiley & Sons, New York
- Bowman S, Lawson D, Basham D, et al. (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* 400:532–538
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94
- Burge CB, Tuschl T, Sharp PS (1999) Splicing precursors to mRNAs by the spliceosomes. In: Gesteland RF, Cech TR, Atkins JF (eds) *The RNA world*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 525–560
- Burset M, Guigo R (1996) Evaluation of gene structure prediction programs. *Genomics* 34:353–367
- Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* 27:3219–3228
- Fink GR (1987) Pseudogenes in yeast? *Cell* 49:5–6
- Fuhrmann M, Oertel W, Hegemann P (1999) A synthetic gene coding for the green fluorescent protein (GFP) is a versatile reporter in *Chlamydomonas reinhardtii*. *Plant J* 19:353–361
- Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Alien J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511
- Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* 11:540–564
- Glöckner G (2000) Large scale sequencing and analysis of AT rich eukaryote genomes. *Curr Genomics* 1:289–299
- Glöckner G, Eichinger L, Szafranski K, Pachebat JA, Bankier AT, Dear PH, Lehmann R, Baumgart C, Parra G, Abril JF, Guigo R, Kumpf K, Consortium DGS, Cox E, Quail MA, Platzer M, Rosenthal A, Noegel AA (2002) Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* 418:79–85
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldman H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science* 274:576, 563–567
- Guigo R (1999) DNA composition, codon usage and exon prediction. In: Bishop M (ed) *Genetic databases*. Academic Press, San Diego, CA, pp 53–80
- Hall N, Pain A, Berriman M, Churcher C, Harris B, Harris D, Mungall K, Bowman S, Atkin R, Baker S, Barron A, Brooks K, Buckee CO, Burrows C, Cherevach I, Chillingworth C, Chillingworth T, Christodoulou Z, Clark L, Clark R, Corton C, Cronin A, Davies R, Davis P, Dear P, Dearden F, Doggett J, Feltwell T, Goble A, Goodhead I, Gwilliam R, Hamlin N, Hance Z, Harper D, Hauser H, Hornsby T, Holroyd S, Horrocks P, Humphray S, Jagels K, James KD, Johnson D, Kerhornou A, Knights A, Konfortov B, Kyes S, Larke N, Lawson D, Lennard N, Line A, Maddison M, McLean J, Mooney P, Moule S, Murphy L, Oliver K, Ormond D, Price C, Quail MA, Rabinowitsch E, Rajandream MA, Rutter S, Rutherford KM, Sanders M, Simmonds M, Seeger K, Sharp S, Smith R, Squares R, Squares S, Stevens K, Taylor K, Tivey A, Unwin L, Whitehead S, Woodward J, Sulston JE, Craig A, Newbold C, Barrell BG (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* 419:527–531
- Hooper PM, Zhang H, Wishart DS (2000) Prediction of genetic structure in eukaryotic DNA using reference point logistic regression and sequence alignment. *Bioinformatics* 16:425–438
- Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* 9:815–824
- Karlin S (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* 9:335–443
- Kriventseva EV, Gelfand MS (1999) Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn* 17:281–228
- Makarov V (2002) Computer programs for eukaryotic gene prediction. *Brief Bioinform* 3:195–199
- Mathe C, Sagot MF, Schiex T, Rouze P (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* 30:4103–117
- Mourier T, Jeffares DC (2003) Eukaryotic intron loss. *Science* 300:1393
- Parra G, Blanco E, Guigo R (2000) GeneID in *Drosophila*. *Genome Res* 10:511–515
- Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* 10:483–501
- Rivero F (2002) mRNA processing in *Dictyostelium*: Sequence requirements for termination and splicing. *Protist* 153:169–176

- Rogic S, Mackworth AK, Ouellette FB (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res* 11:817–832
- Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, Carninci P, Hayashizaki Y, Tomita M (2002) On biased distribution of introns in various eukaryotes. *Gene* 300:89–95
- Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* 59:24–31
- Wood V, Gwilliam R, Rajandream MA, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415:871–880
- Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* 3:698–709