Large Scale Sequencing and Analysis of AT Rich Eukaryote Genomes

Gernot Glöckner*

Institut für Molekulare Biotechnologie e.V., Department of Genome Analysis, PSF 100813, D-07745 Jena, Germany

Abstract: Environmental pressures can direct genomes from a normal to a more or less pronounced imbalance in the base composition. These pressures seem to occur relatively often since genomes with a deviation from a normal base composition are widespread throughout lower eukaryotes. These genomes show altered codon usage and enrichment for the preferred bases in intron and intergenic regions. Techniques designed for large scale sequencing and assembly of genomes with normal base composition will fail with these unusual genomes. Additionally, the currently available analysis tools are mainly suitable for gene finding in genomes with normal base composition.

In recent years some large scale genome analysis projects involving species with a skew directed to a very high AT content were initiated. These projects are encountered with sequencing, assembly, and gap closure problems due to the high AT content. These problems can only be overcome with additional techniques, which partly were developed and used in the ongoing projects. In this review some characteristic aspects of AT rich genomes, the progress of the *Dictyostelium discoideum* and the *Plasmodium falciparum* projects, as well as techniques specifically used for the sequencing of these genomes are highlighted.

ANALYSIS OF GENOMES

In recent years many bacteria, some archaea, and some eukaryotes were selected for genome sequencing and analysis projects. In a second step functional genomics then will provide insights into basic cellular processes [1]. Economical value or expansion of knowledge was of decisive importance for the selection of distinct species. The analysis of genomes of economically interesting species like crop plants, soy bean, etc. may lead to an increase in food production [2]. The knowledge of genomes of pathogens and parasites will probably help to develop novel drugs against defined targets in the pathogeneses. Other species chosen play a key role for the understanding of biological processes and phenomena. The comparison of whole genomes of species from different or consecutive evolutionary stages will elucidate the kind of key genes in evolutionary processes.

Prokaryotic genomes are relatively small, ranging from one to about 5 Mb, and code for less than 1000 to more than 4000 genes [3,4]. The interests in whole genome sequencing and analysis here focus on pathogeneses or species related to pathogeneses. Total sequences of several prokaryaryotes are available (http://www-fp. mcs.anl. gov/~gaasterland/genomes.html) and the number of species sequenced is growing exponentially. The data now available on whole genomes has furthered the knowledge of basic cell functions as well as mechanisms of pathogenity.

Eukaryote genomes are much more complex with genome sizes ranging from about 8 Mb (Ashbya) to some fern species with more than 100,000 Mb [5]. The smallest gene sets are occupied by simply organised species such as some fungi with about 5000 genes, somewhat smaller than that of *S. cerevisiae*. It is estimated that one of the largest gene sets is that of human beings with about 100,000 genes. Several genomes of parasites causing severe and widespread diseases are now being analysed to find new targets for drugs. In addition, the genomes of an increasing number of lower and higher eukaryotes are analysed due to their characteristics as model organisms [6-8].

In recent years many species were proposed to be model organisms. Species selected as models should qualify at least several of the following criteria: i)a key position in evolution ii) pronounced mutants and phenotypes available iii) genetic linkage map exists iv)ease of laboratory handling v) small genome. Many genes in one organism will be only detectable through the comparison with orthologous genes in other model genomes. Moreover,

*Address correspondence to this author at the Institut fur Molekulare Biotechnologie e.V., Abteilung fur Genomanalyse, PSF 100813, D-07745 Jena, Germany; Tel: ++49(0)-3641-656254; Fax: ++49(0)-3641-656255; Email: gernot@imb-jena.de

1389-2029/00 $25.00+.00

functional analysis can be performed on the lowest evolutionary level using the most primitive organism in which the gene of interest exists. So far only two genomes of model organisms, from yeast [9,10]and *C. elegans* [7], are completed but in the next few years many additional genomes will be finished.

On the evolutionary line from the simple, unicellular yeast to the complex, multicellular human species several organisms would accomplish the criteria for model organisms. But despite the large number of species on every evolutionary stage only a few species are proposed to be valuable targets for large scale sequencing efforts. Especially evolutionary branches occupied by lower eukaryotes so far are represented by only a few model species.

In spite of the now acquired routine in large scale sequencing projects there are many technical problems to overcome, if genomes with unusual GC content are to be analysed. But for want of alternatives unusual base composition is no criterion to exclude species from large scale sequencing projects. Thus, parasites as well as model organisms with extremely high AT content are now being analysed. The technical problems accompanied by such projects as well as some common features of genomes with extreme AT skews are discussed.

BASE COMPOSITION BIASES

Patterns of different base composition biases exist in all species. These biases are caused by several different phenomena. Genes expressed at a higher level than average prefer distinct bases at each position of the codon, thus translational selection may dictate codon usage in many organisms [11,12]. Additionally, within prokaryote genomes strand specific AT and GC skews (G-C/G+C and A-T/A+T, respectively) exist. These are probably partly caused by asymmetry in biochemical processes such as DNA replication and repair, and mutation of the nontranscribed strand during transcription [13]. Larger regions with different GC content (isochores) could also be found in bacterial genomes. In some species the GC content varies in a cyclic fashion around the genome [13].

Mammalian genomes are generally composed of different isochores [14-17]. Five isochores were defined in human ranging from about 35 % GC for the L1 isochore to 65 % for the H3 isochore [18,19]. Here the GC rich isochores define gene rich regions of the genome. The medium base composition for the whole human genome is around 50 %. A base composition of 40 to 60 % AT content may be defined as 'normal' base composition which is shared by e.g. the model organisms mouse, Fugu, zebrafish and Drosophila [20,21].

But there are many organisms, prokaryotes and eukaryotes, where the base composition of the whole genome is shifted to extreme GC or AT values due to a variety of environmental pressures. Commonly the base composition skew is designated as mutational bias. The directional mutation theory predicts decreasing mutation frequencies after an initial mutator mutation which may explain the wide variation of the GC-content among species [22]. These pressures superimpose the effects of preferred codons for efficient transcription and imply altered codon usage [23,24] and dinucleotide relative abundance values (the genome signature) [25,26].

A shift of the whole genome to an extreme AT content can be observed in the prokaryote *Borrelia burgdorferi* [27]. The genome of this spirochete also shows a very high strand specific skew. Another unusual feature of this genome is that it consists of a linear chromosome and several linear plasmids. Yet another spirochete, *Treponema pallidum* [28], possesses the same large strand specific skew as *B. burgdorferi*, but has a quite normal base composition and a circular chromosome. Obviously, the different base compositions of the two spirochetes also affect the amino acid composition of their proteins. But the base composition skew is independent of the strand specific skew and only species specific [12].

High mutational biases can also be observed in various lower eukaryote species. GC values in different species reach up to 70% [29]. The highest AT content so far described has *Plasmodium falciparum* with around 80 %. These values represent the upper and lower extreme of the base composition bias in genomes reported so far. In (Table **1**) the genomic features of selected lower eukaryote genomes with very high AT content, and for which a reasonable number of sequence data is available, are summarised. Related species display yet a normal base composition [30]. The genome sizes of the species listed differ in a wide range, they belong all to different phylogenetic branches, and they occupy different biotopes. Due to this diversity it is not easy to define a common mutational pressure for the conversion of balanced to highly biased genomes.

Table 1. Species with AT Rich Genomes

Species	Taxonomy	Chromosome # (haploid)	AT Content	Genome Size	Reference	Comment
Plasmodium falciparum	protists (Apicomplexa)	14	80%	30 Mb	[93] [68]	causes malaria, sequencing of the whole genome, cDNA sequencing project
Tetrahymena thermophila	protists (ciliates)	5	73%	220 Mb	[81]	studies on DNA rearrangement, chromatin assembly etc.,
Dictyostelium discoideum	protists (Dictyosteliida)	6	78%	34 Mb	[62]	model for cell motility cytoskeleton, signaltransduction, sequencing of the whole genome, cDNA sequencing project
Hydra attenuata	cnidarians	15	71%	1.6 Gb	http://xenopus.biochem.uci.edu/facts.html	-
Brugia malayi	nematodes	5	75%	100 Mb	[95]	causes elephantiasis, cDNA sequencing project

THE SPECIES

This work concentrates on the highly biased genomes of *D. discoideum* and *P. falciparum* since considerable results of both sequencing projects are now available. The solutions to problems occurring with the analysis of these AT rich genomes may facilitate the initiation of other sequencing projects involving AT rich species.

The unicellular species *Plasmodium falciparum* belongs to the subgroup of the hemosporidian order within the Apicomplexa, a family in which disease causing species are abundant. Several parasites of this family contain an organelle resembling chloroplasts [31]. Due to phylogenetic relationships of the organellar genome they are thought to be descendants of red algae [32]. The complicated life cycle of *Plasmodium* is predominantly haploid and has obligate stages in human and mosquitoes [33]. The short diploid phase after mating is an absolute requirement for the completion of the life cycle. Malaria originates during the prolonged phases of asexual multiplication of the *Plasmodium* parasites within erythrocytes. Malaria causes 2.7 million deaths annually [34] and many parasite lines are drug resistant [35]. For the design of effective vaccines the knowledge of the whole genome will be very useful. Despite their different codon usage [36] *Plasmodium* species seem to share some regions of synteny [37]. Thus, the evolutionary distance of the *Plasmodium* genomes may not be a hindrance to the understanding of all genomes if one is completely analysed.

P. falciparum has a nuclear genome of about 30 Mb. It is divided between 14 chromosomes, which range in size from 0.7 to 3.5 Mb [38]. The mitochondrial DNA is much less AT rich (68 %) [39]. The other extrachromosomal element is located in a spherical body which is a plastid remnant [40,41]. The AT content of the 35 kb molecule reaches 86.9 % (Acc. No's X95275, X95276).

Dictyostelium discoideum belongs to a diverse yet monophyletic group of organisms, the so-called slime molds. This group (Mycetozoa) includes cellular, acellular and protostelid slime molds. The mycetozoan assemblage is placed within the crown of the eukaryote tree and seems to constitute a sister group to animals and fungi [42]. This placement is consistent with data on other features like physiology or biochemistry [43]. On starvation *D. discoideum* undergoes a developmental life cycle. Individual cells gather to form a slug consisting of thousands of cells [44]. This slug then coordinately migrates and finally builds a fruiting body containing spores which survive awkward environmental conditions, even the ingestion by nematodes [45].

Thus, *D. discoideum* has a unicellular as well as a multicellular life stage. But it is unlikely that this organism represents a direct link between the unicellular yeast and the multicellular *Caenorhabditis elegans* as it was postulated [42,46]. Nevertheless, it can be used as a model organism to study a wide range of characteristics of higher eukaryotes.

The nuclear genome of *D. discoideum* is about 34 Mb in length and distributed over 6 chromosomes ranging from 4.5 Mb to 7 Mb. Additionally, a palindromic linear DNA molecule (inverted repeat) of about 90 kb is contained in the nucleus, which encodes for the rRNA genes [47]. This palindrome is amplified and represents about 20 % of all nucleotides in *D. discoideum.* The nucleotides of the 55 kb mitochondrial genome add up to another 30 % of nucleotide mass so that only 50 % of the nucleotide content of *D. discoideum* represents the real genome. Both extrachromosomal elements could be assembled using contaminating shotgun reads of the genome project (http://dictygenome.bioch.bcm.tmc.edu/extrachromosomal).

In contrast to the genome of *P. falciparum* the genome of *D. discoideum* also contains a family of transposons derived from different origins [48,49]. All transposable elements add up to about 10 % of the genome. Additional sequence data on repetitive elements in *D. discoideum* can be retrieved from http://genome.imb-jena.de/dictyostelium/repeats. A paper describing the features and abundance of complex repetitive elements in the genome is in preparation.

GENETIC AND PHYSICAL MAPS

Assemblies of large genomic fragments could lead to missassembled regions caused by repetitive elements, multigene families, and low complexity regions. The larger the genomic pieces which have to be assembled in one step the more a cross-check for right assembly is needed due to increasing error rates in automated assembly processes. Thus, a prerequisite for genome sequencing projects is the availability of a detailed and highly accurate map in combination with a large-insert clone library of all chromosomes. There are several methods to construct maps of genomes, which also were used for mapping the AT rich genomes.

If a genome is considerable small, single genes can be mapped by the separation of the chromosomes using pulsed field gel electrophoresis (PFGE), subsequent blotting of the gel onto a nylon membrane and hybridisation of the gene probe [50,51].

Since in *D. discoideum* the sexual recombination efficiency is low, parasexuell segregation of heterodiploid strains was initially used to assign genes to specific chromosomes [52,53]. RFLPs of strains harbouring a vector containing rare restriction sites [54] helped to enhance the resolution of the map. To increase the resolution of the chromosome maps a further mapping project using the 'happy mapping' approach [55] was started. With this '*in vitro* meiosis' technique a map of chromosome 6 with a resolution of 8-10 kb will be constructed very fast.

Low resolution restriction maps for the 14 chromosomes of *P. falciparum* were also constructed [56]. Microsatellite markers were constructed and assigned to linkage groups [57]. The assignment of these markers to specific chromosomes was then achieved by the use of PCR on single chromosomes obtained after separation with PFGE (PFG-PCR) [58]. Very recently a genetic map has been constructed using 901 markers [59].

Clone maps of AT rich organisms are based solely on YAC clones due to the inability to construct BAC or PAC libraries. Thus YAC libraries with insert sizes around 100 to 200 kb even with the AT rich DNA were created for *P. falciparum* [60,61] as well as *D. discoideum* [62]. The YAC libraries then were used to build reliable clone maps of each chromosome [56,63] and integrate them with the physical maps. Thus at the beginning of the sequencing project a medium resolution map for both organisms was available.

WHOLE GENOME SHOTGUN

The feasibility of a whole-genome sequencing strategy to obtain the complete nucleotide sequence of an organism was first proven with the free-living bacterium *Haemophilus influenzae* [64]. Meanwhile several microbial genomes have been successfully sequenced with this approach and a lot more are currently in progress. But the largest successfully finished assembly project so far did not exceed 2.5 Mb [65]. At the moment an approach to finish the 165 Mb *Drosophila* genome is in progress [66]. Due to the high AT content of both the genomes of *D. discoideum* and *P. falciparum* the amount of low complexity regions is very high. Long homopolymer runs cause cumulate sequencing

errors in AT stretches and more difficulties and errors in the assembly process. Thus, to reduce the complexity in the assembly process smaller portions of the genome, the chromosomes, can be used for the preparation of shotgun libraries. In both projects this approach was applied. The produced reads then can be assembled to whole chromosomes. Chromosomes up to several Mb can be separated by PFGE but the larger the chromosomes are and the smaller the size differences between them are the less accurate is this separation. Thus due to this incomplete separation cross contamination with clones of other chromosomes occurs in the chromosome specific clone libraries and leads to more difficult assemblies of the raw reads [67]. After the assembly of *P. falciparum* chromosome 3, which is 1,060,106 bp long, 13 % contaminating reads from other chromosomes remained singlets [68]. To test the clone library of chromosome 2 of *D. discoideum* for contamination with DNA from other chromosomes all mapped and sequenced genes of *D. discoideum* were blasted against 10,000 raw reads of the library. Only 50 % of the genes, for which reads were available, were previously mapped to chromosome 2. Considering these results the contamination of the chromosome 2 specific library with other genomic DNA can also be estimated to be about 50 %. This high rate of clones not belonging to chromosome 2 is due to more difficulties in separating the larger chromosomes of *D. discoideum* than that of *P. falciparum* [51]. Thus the reduction of complexity in the assembly process using separated chromosomes as a sequencing resource is only possible if the chromosomes of the organism are not larger than about 7 Mb.

TECHNIQUES AND RESOURCES FOR THE ASSEMBLY OF GENOMES OF AT RICH ORGANISMS

The extreme AT-content makes the genome analysis of both organisms a challenge. The same technical problems arise for both species, and the same solutions for these problems can be used.

No bacterial clone libraries with large insert sizes are available for AT rich organisms. This lack is due to the instability of AT rich DNA larger than about 5 kb if *E. coli* is used as a host. *E. coli* seems to degrade the AT rich inserts maintaining mainly vector sequences and contaminating foreign DNA with normal GC content. Probably only the use of a host vector system specifically designed for AT rich clones would be a solution to this problem.

Additionally, the cloning of DNA fragments skewed to a high AT content is less efficient than the cloning of DNA fragments with 'normal' base composition. Thus, the proportion of clones recovered from a mixture of AT rich DNA with normal GC DNA is shifted to clones containing the DNA with an even base distribution. Thus, in a shotgun approach using large insert bacterial clones the vector sequences with higher GC content would be preferentially cloned. This phenomenon is also a drawback if different isochores or DNA of different types are present in one AT rich organism. This is the case for e.g. *Plasmodium vivax* whose genome is divided into two components of 18 % and 30 % GC, respectively [69]. Additionally, several filarial species like *Brugia malayi* (Table 1) contain an intracellular bacterium with resemblance to rickettsiae [70]. This endosymbiont has also a much higher GC content than the host which leads to excess representation in clone libraries [71].

The finishing of the first two chromosomes of *P. falciparum* has proven that even long stretches of only AT regions can be resolved. But to achieve this several special methods have to be used.

Clones representing intergenic AT rich clones are underrepresented in small insert libraries. The ability to get high quality sequences in those regions is also limited. A first analysis of raw reads of the *Dictyostelium* genome showed that sequences of the intergenic regions are underrepresented by a factor of 2. This drawback could be partly circumvented by increasing the number of reads produced to a 15 fold coverage as for *Plasmodium* Chr3. This approach increases the costs of the sequencing project. The automated assembly is also hindered by large intergenic regions with low complexity. The use of pUC18 instead of m13mp18 as sequencing vector enables the generation of sequence information from both ends of the insert. This linked sequence information can then be used to order contigs along the chromosome according to clones spanning sequencing gaps.

Besides the standard techniques used for gap closure for AT rich organisms additional procedures have to be applied. For the closure of long gaps with very high AT content a transposon insertion technique can be used [72,73]. An artificial transposon is integrated at various sites in plasmids spanning such sequencing gaps, leading to new starting points for sequencing. This method was successfully used for the closure of some gaps in the finished *Plasmodium* chromosomes. Another approach to the closure of large gaps is the use of

small insert libraries with m13mp18 as vectors derived from one single plasmid [68].

Additionally, after assembly different methods should be used to check for errors in the assembly. A direct method is the generation of a restriction map based on direct measurement of the fragment lengths using a microscope. This optical map is highly accurate [74] and even from large DNA molecules a restriction fragment map can be constructed [75-77].

The skimming of yeast artificial chromosome (YAC) clones is another cross-check method used [68]. Here a shotgun library of a YAC clone is constructed and sequences of only a few clones not sufficient for the assembly of the whole YAC are generated. This sequence information is then used to bin sequences produced from the whole genome or chromosome shotgun library. Yet the separation of the YAC from the other yeast chromosomes is a time consuming and often ineffective procedure. Typically, the YAC preparations are contaminated with a certain amount of other yeast chromosomes which are then preferentially cloned in bacterial vectors. Thus the yeast DNA contamination in a shotgun library of a YAC clone leads to an inefficient skimming procedure. A method which uses the positional information of YAC clones directly is thus preferable. A procedure for the determination of unknown DNA segments adjacent to known sequences [78] was recently adapted to the special conditions of AT rich genomes [79]. With this method it may be possible to generate end sequences of YAC clones as fast and inexpensive as for bacterial clones. Thus even without a YAC map generated with standard hybridisation and restriction fragment length comparisons it will be possible to use the YAC resources for the correct assembly of AT rich genomes.

GENOMIC STRUCTURE, GC DISTRIBUTION, AND GENE DETECTION

Analysis of gene structures in an AT rich environment revealed that long (dA).(dT) tracts are found preferentially in introns and gene flanking regions [80]. Thus non coding regions make a major contribution to high AT levels in genomes [81]. Coding regions have to contain a certain amount of G and C due to theoretical limitations of the base composition necessary to direct insertion of all 20 amino acids [23]. It seems to be impossible to achieve a higher mean AT level than 80 % as in *P. falciparum* in genomes with comparable gene density. Extension of intergenic regions and introns in less compact genomes would yet lead to slightly higher AT values.

Most genes in both organisms contain no or few and small introns and the mean density of genes is around 1 gene/4.5 kb. This value is only an estimation for *D. discoideum* deduced from a few genomic sequences since currently no large genomic DNA fragments are available.

Many genes of both organisms contain low complexity regions with AT rich codons. These regions are translated and result in nonglobular homopolymer runs. The homopolymer runs of *P. falciparum* include asparagine, lysine, and glutamic acid as was analysed on the finished two chromosomes. Preliminary data from *D. discoideum* genes show a comparable amount and distribution but slightly longer stretches of those homopolymer runs. These expansions of A/T rich codons thus seem to be a common feature of genes of A/T rich organisms. On the other hand comparable extensions of homopolymer runs could be observed in GC rich genes [82]. Thus, every pressure leading to extreme base composition skews may favour long homopolymeric amino acid repeats in proteins.

(Fig. **1)** shows a 23 kb genomic region of *Dictyostelium*. At a first glance the regions with coding potential are easily detectable. Roughly every segment with a GC content above 22 % is coding. The gene models depicted here span regions of variable GC levels, have a long ORF, and at least one of the exons of each gene model gave a blast hit against the genpept database. Thus, there may be more exons but not more genes present in this region. This DNA region also shows that the exons forming a gene model are not clearly separated from exons of other gene models. The lengths of intergenic regions are variable and may contain a considerable amount of G and C bases. On the other hand introns may be void of G and C and exhibit a minimum in the GC curve. These features make it difficult to assign single exons to a distinct gene model.

The different GC content of coding and non-coding regions is the basis for melting curves, which differ accordingly. The experimental determination of melting curves for *Dictyostelium* DNA was replaced by an analysis of the respective melting temperatures *in silico* [83]. The MELTSIM software used for this purpose [84] can be retrieved from http://www.uml.edu/Dept/Chem/Bioinfo/-Apps/MELTSIM.

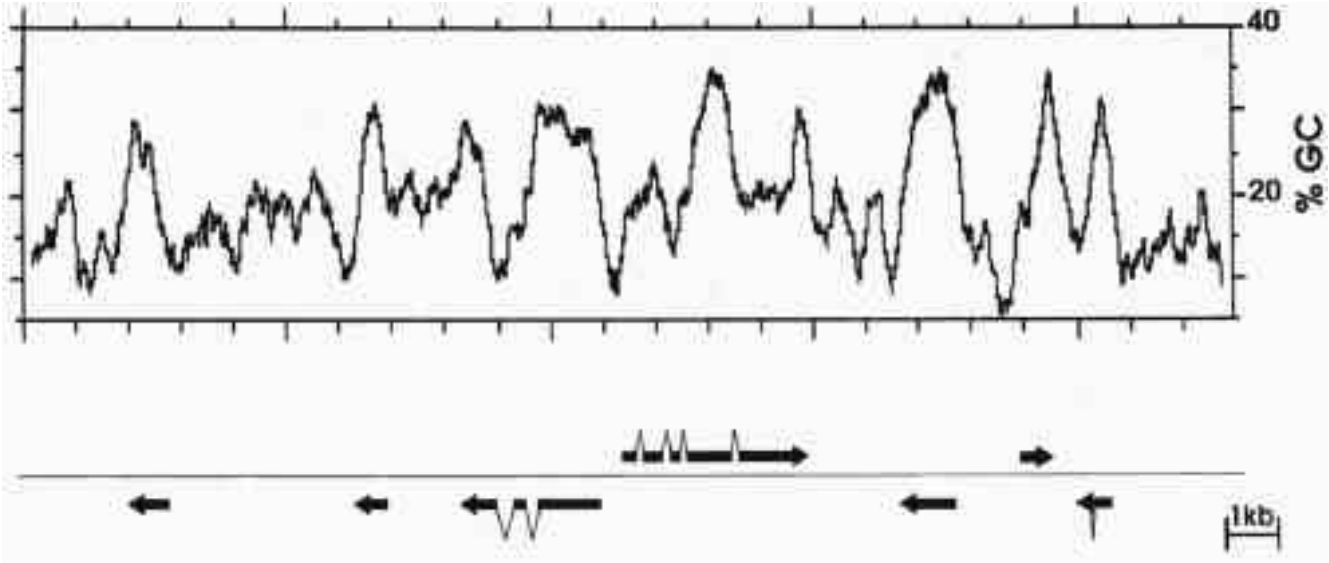

Fig. (1). GC-plot of a 23 kb segment (81 % AT) of *Dictyostelium* genomic DNA calculated with a window size of 400 and a step of 1. Preliminary gene models derived from similarities to known genes are depicted as arrows indicating the transcription direction.

The analysis of long range correlations in the genome is also a powerful tool for the definition of genes. A recurrence plot from a 47 kb sequence in *P. falciparum* shows that intergenic regions are composed of highly recurrent oligonucleotides [85]. This feature is more pronounced in AT rich genomes than in others with normal base composition.

Mung bean nuclease is able to cut double stranded DNA depending on AT content and reaction temperature [86]. Thus in AT rich organisms the intergenic regions and long intron sequences are preferentially cut due to their lower melting temperature. The clones of the resulting fragments constitute the gene complement of the organism [56]. This approach could be especially useful for the analysis of large AT rich genomes like that of Hydra where the sequencing of the whole genome seems to be to time consuming and expensive.

The most commonly used programmes for gene detection which also give gene models (Grail [87] and Genscan [88]) are not yet adapted to the unusual conditions of AT rich genomes. Since the detection of regions with coding potential in AT rich genomes seems to be easy the greatest efforts should be invested for the building of reliable gene models. One very important tool for this purpose is the avaibility of ESTs [89]. A considerable number of ESTs from different life stages are being produced [90,91] for both organisms facilitating thus the gene detection. Orthologues and the confirmation by RT-PCR can then be used to confirm the gene models.

CURRENT STATUS OF THE PROJECTS

Since the nuclear chromosomes of *P. falciparum* are polymorphic varying significantly between clones [56] a reference clone had to be chosen for the genome project. Clone 3D7 has been selected because it grows well in vitro, has not suffered the loss of known functions, and can be used for genetic crosses[92]. Three centres share the responsibility for sequencing the genome of *P. falciparum.* The genome is split on a chromosome by chromosome basis between The Sanger Centre (U.K.), The Institute for Genomic Research /Naval Medical Research Institute (U.S.A.), and Stanford University (U.S.A.). Currently two of the smallest chromosomes (2 and 3) are sequenced and analysed[68,93]. The production of shotgun reads and the gap closure is well underway for the remaining 12 chromosomes (http://www.sanger.-ac.uk/Projects/P_falciparum/who&what.shtml).

All isolates and clones of *D. discoideum* vary slightly in their physiological responses [94]. For sequencing the clone AX4 was selected since for this strain a genetic and physical map was constructed [63], which will be needed for the

assembly check of the chromosomes. The chromosomes are being isolated at Princeton University (U.S.A.), then the libraries for all 6 chromosomes are being constructed at the Sanger Centre. The sequencing efforts here are also shared among three centres: The Genome Sequencing Centre at the IMB in Jena in collaboration with the University of Cologne (D), the Sanger Centre (U.K.), and the Baylor College of Medicine (U.S.A.). Additional YAC skims of chromosome 6 are delivered by the Institute Pasteur (F). Currently the production of shotgun reads of the chromosomes 1 and 2 (Jena) and 6 (Baylor College and Sanger Centre) is underway. Sequencing of the remaining three chromosomes (3, 4, 5) will be started in the year 2000. The first assembled chromosome will be available in mid 2000. All sequencing data can be retrieved from http://genome.imb-jena.de/dictyostelium or http://www.uni-koeln.de/dictyostelium.

Despite the difficulties in large scale genome sequencing and analysis of AT rich genomes considerable progress in handling such projects has been made during the last few years. When the complete genomes of both organisms will be available a careful comparative analysis may reveal the driving forces for high AT skews.

ACKNOWLEDGMENT

I thank P. Fröhlich for critical reading of the manuscript. Research by the author is supported by the DFG.

REFERENCES

[1] Fields, S.; Kohara, Y. and Lockhart, D. J. (1999) *Proc Natl Acad Sci U S A*, **96**(16), 8825-6.

[2] Kasha, K. J. (1999) *Genome*, **42**(4), 642-5.

[3] Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A. M.; Alloni, G.; Azevedo, V.; Bertero, M. G.; Bessieres, P.; Bolotin, A.; Borchert, S.; Borriss, R.; Boursier, L.; Brans, A.; Braun, M.; Brignell, S. C.; Bron, S.; Brouillet, S.; Bruschi, C. V.; Caldwell, B.; Capuano, V.; Carter, N. M.; Choi, S. K.; Codani, J. J.; Connerton, I. F.; Danchin, A. and *et al.* (1997) *Nature*, **390**(6657), 249-56.

[4] Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Clayton, R. A.; Fleischmann, R. D.; Bult, C. J.; Kerlavage, A. R.; Sutton, G.; Kelley, J. M. and *et al.* (1995) *Science*, **270**(5235), 397-403.

[5] Vinogradov, A. E. (1999) *J Mol Evol.*, **49**(3), 376-84.

[6] Botstein, D.; Chervitz, S. A. and Cherry, J. M. (1997) *Science*, **277**(5330), 1259-60.

[7] Kitano, H.; Hamahashi, S. and Luke, S. (1998) *Artif Life*, **4**(2), 141-156.

[8] Haffter, P. and Nusslein-Volhard, C. (1996) *Int J Dev Biol.*, **40**(1), 221-7.

[9] Hudson, J. R.; Jr.; Dawson, E. P.; Rushing, K. L.; Jackson, C. H.; Lockshon, D.; Conover, D.; Lanciault, C.; Harris, J. R.; Simmons, S. J.; Rothstein, R. and Fields, S. (1997) *Genome Res.*, **7**(12), 1169-73.

[10] Johnston, M. (1996) *Curr. Biol.*, **6**(5), 500-3.

[11] Pan, A.; Dutta, C. and Das, J. (1998) *Gene.*, **215**(2), 405-13.

[12] Lafay, B.; Lloyd, A. T.; McLean, M. J.; Devine, K. M.; Sharp, P. M. and Wolfe, K. H. (1999) *Nucleic Acids Res.*, **27**(7), 1642-9.

[13] McLean, M. J.; Wolfe, K. H. and Devine, K. M. (1998) *J. Mol. Evol.*, **47**(6), 691-6.

[14] Clay, O.; Caccio, S.; Zoubak, S.; Mouchiroud, D. and Bernardi, G. (1996) *Mol. Phylogenet. Evol.*, **5**(1), 2-12.

[15] Mouchiroud, D.; Fichant, G. and Bernardi, G. (1987) *J. Mol. Evol.*, **26**(3), 198-204.

[16] Bernardi, G. (1993) *Mol. Biol. Evol.*, **10**(1), 186-204.

[17] Robinson, M.; Gautier, C. and Mouchiroud, D. (1997) *Mol. Biol. Evol.*, **14**(8), 823-8.

[18] Sabeur, G.; Macaya, G.; Kadi, F. and Bernardi, G. (1993) *J. Mol. Evol.*, **37**(2), 93-108.

[19] Saccone, S.; Caccio, S.; Kusuda, J.; Andreozzi, L. and Bernardi, G. (1996) *Gene.*, **174**(1), 85-94.

[20] Cross, S.; Kovarik, P.; Schmidtke, J. and Bird, A. (1991) *Nucleic Acids Res.*, **19**(7), 1469-74.

[21] Moreau, J.; Kejzlarova-Lepesant, J.; Brock, H.; Lepesant, J. A. and Scherrer, K. (1985) *Mol. Gen. Genet.*, **199**(3), 357-64.

[22] Sueoka, N. (1993) *J. Mol. Evol.*, **37**(2), 137-53.

[23] Winkler, H. H. and Wood, D. O. (1988) *Biochimie*, **70**(8), 977-86.

[24] Nakamura, Y.; Gojobori, T. and Ikemura, T. (1999) *Nucleic Acids Res.*, **27**(1), 292.

[25] Karlin, S. (1998) *Curr. Opin. Microbiol.*, **1**(5), 598-610.

[26] Karlin, S.; Campbell, A. M. and Mrazek, J. (1998) *Annu. Rev. Genet.*, **32**, 185-225.

[27] Fraser, C. M.; Casjens, S.; Huang, W. M.; Sutton, G. G.; Clayton, R.; Lathigra, R.; White, O.; Ketchum, K. A.; Dodson, R.; Hickey, E. K.; Gwinn, M.; Dougherty, B.; Tomb, J. F.; Fleischmann, R. D.; Richardson, D.; Peterson, J.; Kerlavage, A. R.; Quackenbush, J.; Salzberg, S.; Hanson, M.; van Vugt, R.; Palmer, N.; Adams, M. D.; Gocayne, J.; Venter, J. C. and *et al.* (1997) *Nature*, **390**(6660), 580-6.

[28] Fraser, C. M.; Norris, S. J.; Weinstock, G. M.; White, O.; Sutton, G. G.; Dodson, R.; Gwinn, M.; Hickey, E. K.; Clayton, R.; Ketchum, K. A.; Sodergren, E.; Hardham, J. M.; McLeod, M. P.; Salzberg, S.; Peterson, J.; Khalak, H.; Richardson, D.; Howell, J. K.; Chidambaram, M.; Utterback, T.; McDonald, L.; Artiach, P.; Bowman, C.; Cotton, M. D.; Venter, J. C. and *et al.* (1998) *Science*, **281**(5375), 375-88.

[29] Merchant, S.; Hill, K.; Kim, J. H.; Thompson, J.; Zaitlin, D. and Bogorad, L. (1990) *J. Biol. Chem.*, **265**(21), 12372-9.

[30] Wilson, R. K. (1999) *Trends Genet.*, **15**(2), 51-8.

[31] Lang-Unnasch, N.; Reith, M. E.; Munholland, J. and Barta, J. R. (1998) *Int. J. Parasitol.*, **28**(11), 1743-54.

[32] Blanchard, J. L. and Hicks, J. S. (1999) *J. Eukaryot. Microbiol.*, **46**(4), 367-75.

[33] Arnot, D. E. and Gull, K. (1998) *Ann. Trop. Med. Parasitol.*, **92**(4), 361-5.

[34] Nussenzweig, R. S. and Long, C. A. (1994) *Science*, **265**(5177), 1381-3.

[35] Serrano, A. E.; Robinson, B. L.; Peters, W. and Trujillo-Nevarez, K. (1999) *Exp. Parasitol.*, **91**(1), 93-6.

[36] Chen, N. and Cheng, Q. (1999) *Int. J. Parasitol.*, **29**(3), 445-9.

[37] Carlton, J. M.; Vinkenoog, R.; Waters, A. P. and Walliker, D. (1998) *Mol. Biochem. Parasitol.*, **93**(2), 285-94.

[38] Walker-Jonah, A.; Dolan, S. A.; Gwadz, R. W.; Panton, L. J. and Wellems, T. E. (1992) *Mol. Biochem. Parasitol.*, **51**(2), 313-20.

[39] Feagin, J. E. (1992) *Mol. Biochem. Parasitol.*, **52**(1), 145-8.

[40] Feagin, J. E.; Werner, E.; Gardner, M. J.; Williamson, D. H. and Wilson, R. J. (1992) *Nucleic Acids Res.*, **20**(4), 879-87.

[41] Wilson, R. J.; Denny, P. W.; Preiser, P. R.; Rangachari, K.; Roberts, K.; Roy, A.; Whyte, A.; Strath, M.; Moore, D. J.; Moore, P. W. and Williamson, D. H. (1996) *J. Mol. Biol.*, **261**(2), 155-72.

[42] Baldauf, S. L. and Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci., USA*, **94**(22), 12007-12.

[43] Kay, R. R. (1994) *Curr. Opin. Genet. Dev.*, **4**(5), 637-41.

[44] Parent, C. A. and Devreotes, P. N. (1996) *Annu. Rev. Biochem.*, **65**, 411-40.

[45] Kessin, R. H.; Gundersen, G. G.; Zaydfudim, V. and Grimson, M. (1996) *Proc. Natl. Acad. Sci. USA*, **93**(10), 4857-61.

[46] Loomis, W. F. and Smith, D. W. (1995) *Experientia.*, **51**(12), 1110-5.

[47] Parish, R. W.; Banz, E. and Ness, P. J. (1986) *Nucleic Acids Res.*, **14**(5), 2089-107.

[48] Winckler, T. (1998) *Cell. Mol. Life. Sci.*, **54**(5), 383-93.

[49] Szafranski, K.; Glöckner, G.; Dingermann, T.; Noegel, A. A.; Eichinger, L.; Rosenthal, A. and Winckler, T. (1999) *Mol. Gen. Genet.*, (in press).

[50] Hernandez-Rivas, R. and Scherf, A. (1997) *Mem Inst Oswaldo Cruz*, **92**(6), 815-9.

[51] Cox, E. C.; Vocke, C. D.; Walter, S.; Gregg, K. Y. and Bain, E. S. (1990) *Proc. Natl. Acad. Sci. USA*, **87**(21), 8247-51.

[52] Rothman, F. G. and Alexander, E. T. (1975) *Genetics*, **80**(4), 715-31.

[53] Welker, D. L. and Williams, K. L. (1982) *Genetics*, **102**(4), 691-710.

[54] Kuspa, A. and Loomis, W. F. (1994) *Genetics*, **138**(3), 665-74.

[55] Dear, P. H. and Cook, P. R. (1993) *Nucleic Acids Res.*, **21**(1), 13-20.

[56] Dame, J. B.; Arnot, D. E.; Bourke, P. F.; Chakrabarti, D.; Christodoulou, Z.; Coppel, R. L.; Cowman, A. F.; Craig, A. G.; Fischer, K.; Foster, J.; Goodman, N.; Hinterberg, K.; Holder, A. A.; Holt, D. C.; Kemp, D. J.; Lanzer, M.; Lim, A.; Newbold, C. I.; Ravetch, J. V.; Reddy, G. R.; Rubio, J.; Schuster, S. M.; Su, X. Z.; Thompson, J. K.; Werner, E. B. and *et al.* (1996) *Mol. Biochem. Parasitol.*, **79**(1), 1-12.

[57] Su, X. and Wellems, T. E. (1996) *Genomics*, **33**(3), 430-44.

[58] Su, X. Z. and Wellems, T. E. (1999) *Exp. Parasitol.*, **91**(4), 367-9.

[59] Su, X.; Ferdig, M. T.; Huang, Y.; Huynh, C. Q.; Liu, A.; You, J.; Wootton, J. C. and Wellems, T. E. (1999) *Science*, **286**(5443), 1351-3.

[60] Triglia, T. and Kemp, D. J. (1991) *Mol. Biochem. Parasitol.*, **44**(2), 207-11.

[61] Rubio, J. P.; Triglia, T.; Kemp, D. J.; de Bruin, D.; Ravetch, J. V. and Cowman, A. F. (1995) *Genomics*, **26**(2), 192-8.

[62] Kuspa, A.; Maghakian, D.; Bergesch, P. and Loomis, W. F. (1992) *Genomics*, **13**(1), 49-61.

[63] Kuspa, A. and Loomis, W. F. (1996) *Proc. Natl. Acad. Sci. USA*, **93**(11), 5562-6.

[64] Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M. and *et al.* (1995) *Science*, **269**(5223), 496-512.

[65] Klenk, H. P.; Clayton, R. A.; Tomb, J. F.; White, O.; Nelson, K. E.; Ketchum, K. A.; Dodson, R. J.; Gwinn, M.; Hickey, E. K.; Peterson, J. D.; Richardson, D. L.; Kerlavage, A. R.; Graham, D. E.; Kyrpides, N. C.; Fleischmann, R. D.; Quackenbush, J.; Lee, N. H.; Sutton, G. G.; Gill, S.; Kirkness, E. F.; Dougherty, B. A.; McKenney, K.; Adams, M. D.; Loftus, B.; Venter, J. C. and *et al.* (1997) *Nature*, **390**(6658), 364-70.

[66] Little, P. (1999) *Nature*, **402**, 468.

[67] Lawson, D. (1999) *Parasitology*, **118** Suppl, S15-8.

[68] Bowman, S.; Lawson, D.; Basham, D.; Brown, D.; Chillingworth, T.; Churcher, C. M.; Craig, A.; Davies, R. M.; Devlin, K.; Feltwell, T.; Gentles, S.; Gwilliam, R.; Hamlin, N.; Harris, D.; Holroyd, S.; Hornsby, T.; Horrocks, P.; Jagels, K.; Jassal, B.; Kyes, S.; McLean, J.; Moule, S.; Mungall, K.; Murphy, L.; Barrell, B. G. and *et al.* (1999) *Nature*, **400**(6744), 532-8.

[69] Camargo, A. A.; Fischer, K.; Lanzer, M. and del Portillo, H. A. (1997) *Genomics*, **42**(3), 467-73.

[70] Bandi, C.; Anderson, T. J.; Genchi, C. and Blaxter, M. L. (1998) *Proc. R. Soc. Lond. B. Biol. Sci.*, **265**(1413), 2407-13.

[71] Ash, C. (1999) *Trends Microbiol.*, **7**(1), 10-2.

[72] Devine, S. E. and Boeke, J. D. (1994) *Nucleic Acids Res.*, **22**(18), 3765-72.

[73] Devine, S. E.; Chissoe, S. L.; Eby, Y.; Wilson, R. K. and Boeke, J. D. (1997) *Genome Res.*, **7**(5), 551-63.

[74] Anantharaman, T. S.; Mishra, B. and Schwartz, D. C. (1997) *J. Comput. Biol.*, **4**(2), 91-118.

[75] Jing, J.; Lai, Z.; Aston, C.; Lin, J.; Carucci, D. J.; Gardner, M. J.; Mishra, B.; Anantharaman, T. S.; Tettelin, H.; Cummings, L. M.; Hoffman, S. L.; Venter, J. C. and Schwartz, D. C. (1999) *Genome Res.*, **9**(2), 175-81.

[76] Jing, J.; Reed, J.; Huang, J.; Hu, X.; Clarke, V.; Edington, J.; Housman, D.; Anantharaman, T. S.; Huff, E. J.; Mishra, B.; Porter, B.; Shenker, A.; Wolfson, E.; Hiort, C.; Kantor, R.; Aston, C. and Schwartz, D. C. (1998) *Proc. Natl. Acad. Sci. USA*, **95**(14), 8046-51.

[77] Lin, J.; Qi, R.; Aston, C.; Jing, J.; Anantharaman, T. S.; Mishra, B.; White, O.; Daly, M. J.; Minton, K. W.; Venter, J. C. and Schwartz, D. C. (1999) *Science*, **285**(5433), 1558-62.

[78] Triglia, T.; Peterson, M. G. and Kemp, D. J. (1988) *Nucleic Acids Res.*, **16**(16), 8186.

[79] Foster, J. M.; Christodoulou, Z.; Cowan, G. M. and Newbold, C. I. (1999) *Biotechniques*, **27**(2), 240, 244, 246.

[80] Marx, K. A.; Hess, S. T. and Blake, R. D. (1994) *J. Biomol. Struct. Dyn.*, **12**(1), 235-46.

[81] Wuitschick, J. D. and Karrer, K. M. (1999) *J. Eukaryot. Microbiol.*, **46**(3), 239-47.

[82] Sumiyama, K.; Washio-Watanabe, K.; Saitou, N.; Hayakawa, T. and Ueda, S. (1996) *J. Mol. Evol.*, **43**(3), 170-8.

[83] Marx, K. A.; Assil, I. Q.; Bizzaro, J. W. and Blake, R. D. (1998) *J. Biomol. Struct. Dyn.*, **16**(2), 329-39.

[84] Blake, R. D.; Bizzaro, J. W.; Blake, J. D.; Day, G. R.; Delcourt, S. G.; Knowles, J.; Marx, K. A. and SantaLucia, J.; Jr. (1999) *Bioinformatics*, **15**(5), 370-5.

[85] Frontali, C. and Pizzi, E. (1999) *Gene*, **232**(1), 87-95.

[86] McCutchan, T. F.; Hansen, J. L.; Dame, J. B. and Mullins, J. A. (1984) *Science*, **225**(4662), 625-8.

[87] Uberbacher, E. C.; Xu, Y. and Mural, R. J. (1996) *Methods Enzymol.*, **266**, 259-81.

[88] Burge, C. and Karlin, S. (1997) *J. Mol. Biol.*, **268**(1), 78-94.

[89] Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F. and *et al.* (1991) *Science*, **252**(5013), 1651-6.

[90] Morio, T.; Urushihara, H.; Saito, T.; Ugawa, Y.; Mizuno, H.; Yoshida, M.; Yoshino, R.; Mitra, B. N.; Pi, M.; Sato, T.; Takemoto, K.; Yasukawa, H.; Williams, J.; Maeda, M.; Takeuchi, I.; Ochiai, H. and Tanaka, Y. (1998) *DNA Res.*, **5**(6), 335-40.

[91] Wellems, T. E.; Su, X. Z.; Ferdig, M. and Fidock, D. A. (1999) *Curr. Opin. Microbiol.*, **2**(4), 415-9.

[92] Walliker, D.; Quakyi, I. A.; Wellems, T. E.; McCutchan, T. F.; Szarfman, A.; London, W. T.; Corcoran, L. M.; Burkot, T. R. and Carter, R. (1987) *Science*, **236**(4809), 1661-6.

[93] Gardner, M. J.; Tettelin, H.; Carucci, D. J.; Cummings, L. M.; Aravind, L.; Koonin, E. V.; Shallom, S.; Mason, T.; Yu, K.; Fujii, C.; Pederson, J.; Shen, K.; Jing, J.; Aston, C.; Lai, Z.; Schwartz, D. C.; Pertea, M.; Salzberg, S.; Zhou, L.; Sutton, G. G.; Clayton, R.; White, O.; Smith, H. O.; Fraser, C. M.; Hoffman, S. L. and *et al.* (1998) *Science*, **282**(5391), 1126-32.

[94] Kellerman, K. A. and McNally, J. G. (1999) *Dev Biol.*, **208**(2), 416-29.

[95] Rothstein, N.; Stoller, T. J. and Rajan, T. V. (1988) *Parasitology*, **97**(Pt 1), 75-9.