SCIENTIA YUGOSLAVICA 1990 16(1-2)97-107

Conference paper
UDC 575,113
BSO 413,70-160,30

PROSPECTS FOR A MINIATURIZED, SIMPLIFIED AND FRUGAL HUMAN GENOME PROJECT*

*Radoje Drmanc*** *and Radomir Crkvenjakov*
Genetic Engineering Center, po box 794, 11000 Belgrade, Yugoslavia

Primljeno: 1990-01-15

INTRODUCTION

The knowledge about parts or entire genomes on the level of primary structure as well as the possibility of following inheritance using this information are being increasingly recognized as conditions for more efficient and faster study of biological processes. Some experimental investigations will be replaced by computer research on obtained sequences. It might turn out that some biological phenomena, for example certain evolutionary processes, will be accessible fro study only through the analysis of genome sequences. We are concerned here with the question of haw the application of computers can aid us in compiling the sequence itself.

Two areas are recongizable in which there is an increase in organized efforts to find methodological solutions which would facilitate the determination of primary genetic information. One concerns the detection of a large number of mapped polymorphic sites in the fenomic DNA of individuals, families or populations. This requires the compilation of a defined part of genomic sequence which represents a specific genomic subentity. We propose the name GENOGRAM for thic information.

The other project deals with the determination of the entire sequence of human and other genomes. In the extreme, it might mean the determination of the sequences of the genomes of most species of interes and in sufficient numbers of individuals in each species.

Technically, the first project uses two approaches: the detection of a polymorphic site is based either on the restriction enzyme specificity (RFLP)[1] or hybridization specificitu of oligonucleotide probes[2]. The oligonucleotide probe based approach has considerable advantages in the case of following a large number of sites since it does not require the determination of DNA fragment length and is adapted for use in amplified targets (PCR)[3], oligo-ligation assay[4] or amplified ligationhybridization reaction[5]. One needs to determine the polyporphic status of 5000-10000 sites to obtain individual human genetic maps with a 1 cm resolution.

The second project is envisioned as multiphase physical mapping with the final goal of determining the human genomic sequence[6,7]. Physical mapping can also be accomplished either by sequence recognition by restriction enzymes and measurement of the ensuing fragment lengths[8,9,10] or by the determination of the contents of short sequences by hybridization with oligonucleotides[11,12]. Present methods of sequencing also use the measurement of lengths of fragments and thus determine the sequence[13,14].

Two further approaches for determining genome sequences are being considered as well. One is based on the sequential removal of single nucleotides from one end of an individual DNA fragment and their subsequent detection[6, 7, 15]. The second supposes direst experimental reading of the sequence of a DNA molecule under a powerful electron tunneling microscope[6, 7, 16]. Finally, a theory of an opproach has been developed in which the sequence is not directly determined in an experiment, but the contents of very short oligonucleotide sequences are found instead. Then these data are transformed to sequence information by extencive computation[17]. We have called this approach sequencing by hybridiation (SBH). Presently, the only realistic way for determining the contents of oligo sequences is by oligonucleotide hybridization of the same kind used in the other methods described.

In both projects, one operates with huge number of samples, except possibly in electron tunneling microscopy. Depending on the size and number of the genomes, the number of samples is typically between 10^6 and 10^7. Since each sample is subjected to one or more identical or similar reactions, there is a problem of ways (and speed) of performing repetitive operations as well as a problem of

* Informational approach for the determination of the complete or partial primary structures of complex genomes; requirements and the concept of a workable methodology

** Present address: ICRF, PO Box 123 Lincoln's Inn Fields, London WC2A 3PX, UK

ways and speed of gathering experimental information and storing it in computer memory. Obviously, the solution for the first problem is a robotized process[6] and for the second image analysis [18,19].

Here we define an »Informational approach« for the study of primary genome structure, analyze its characteristics and requirement as well as provide conceptual solutions for its major technological components. These solutions allow the substitution of a massive robotized process using addressed samples by a miniaturized process with randomized samples. This potentially leads to radical technological improvements for the genome projects. Basically, the idea is to use a mixture of discrete recognizable particles, each of which represents an individual sample.

A. THEORETICAL-INFORMATIONAL ANALYSIS

A1. *Requirements for sequencing 10 genomes in a year*

Complex genomes contain several billion base pairs. The use of overlapped DNA fragments and fewfold resequencing is unavoidable, if an acceptable level of accuracy is to be attained. Therefore, one concludes that at least 10^{10} discrete data bits need to be read from experiment into a computer momory to obtain the sequence of a mammalian genome.

Interestingly, for GENOGRAMs one needs a similar number of data bits as well. 10^4 genomic sites or DNA fragment need to be scored to obtain a 1 cm genetic map. If all the genes are scored, the number increases to 10^5. Estimating the requirements of the molecular genetics of the future, at least 1000 individual GENOGRAMS would need to be compiled yearly and the demand might eventually rise to 10^5 for as many patients and individuals in populational studies. Therefore, extensive populational and diagnostic investigations based on human genome sequences will soon reach genome sequencing itself in required experimental work.

One can assume that in the future,the time for determining the sequence of a complex genome cannot be longer than a year. If seconds are used (one year has $3{,}1536 \times 10^7$ seconds), all experimental bits schould be aequired in 10^6 to 10^7 seconds. If one chooses 10^6 seconds (12 days) as a very favorable time span, one should obtain one part in a million of the necessary data every second. This expressed as a 10^{10} bp requirement for complex genomes, establishes the speed for data acquisition at 10000 bp/sec. This estimate should not include just the speed of sequencing operations, but also the time required for sample preparation. One should keep in mind that 12 days from the preceeding argument would be in reality 5-10 times longer because of down time for preparation, control etc.

A2 *Maximizing efficiency by exploiting the potential of IMAGE ANALYSIS as a nonlimiting step*

From the efficiency standpoint, the best way to gather data bits in parallel is to present experimental images in the form of complex matrices, for instance dot blot autoradiographs, sequencing gel or films, etc. The speediest reading of the data is provided by computer aided IMAGE ANALYSIS, since it does not require physical juxtapositioning of the detector and the individual sample in the matrix. In this process, the entire image is projected on the two dimensional array of a large number of photodetectors, pixels each of which independently inputs into the computer memory. From the processing of such inputs (informational work) the computer produces usable data. The only requirement is that the objects in planar image cover an area larger than the individual pixels when projected on the image analyser.

Based on the accomplishments and possibilities of electronics, one can assume that IMAGE ANALYSIS can be very fast and doable in a miniaturized set up[19]. From this assumption, a very important expectation can be derived for genome sequencing strategies: »Reading« of experimental images will represent a nonlimiting part of time and technical requirements in sequence determination, irrespective of the approach. In other words, the efficiency of different sequencing strategies will be directly proportional to the time and technical requirements for obtaining specific experimental images. If this is the case, the number of experimental data bits required by a given approach is not the direct measure of its efficiency. Quite the opposite might be true. Due to expected throughput of IMAGE ANALYSIS, the approaches requiring a larger number of bits might be more efficent than the ones requiring less, but for which the sume total of experimental images are obtained with less material and time investment. We propose the following efficiency measure for genome sequencing methods: the fraction of necessary data for sequence generation at a given accuracy which experiment can present for IMAGE ANALYSIS per unit time.

A3. *Informational approach and its characteristics*

There is a consensus in scientific circles that the existing DNA sequencing methods should be improved and new ones developed in order that the determination of primary structure of complex genomes, or at least human genome, can be achieved[6]. To this end, one can pose the question what is the number of possible ways for determining the sequence of a linear array of a finite number of different but recurring elements. A supplementary question is whether solutions for this informational problem, when applied to existing or proposed sequencing methods, suggest that methods based on new principles are required, or are only the variants of known methods possible. This problem can be resolved by turning it around. The number of informationally different ways for the determination of sequence is equal to the number of principally different ways in which a linear array can be recorded in data bits whose order is immaterial. In other words, how many types of data exist from

whose randomly ordered recording a sequence can be generated? We cannot think of but two types of such data: the first recording principle in which unit data bits are composed from binary designations element/position and the second in which unit data represent integration of binary data, i.e. short subsequences of the starting linear array. This latter way, which uses segments whose position is not specified, can be more precisely defined as a recording of the minimal number of the shortest segments from which a starting sequence can be generated as the unique outcome. One can analyse the question of what repercussions the defined informational basis of the two possible approaches have on the experimental determination of the sequence of DNA fragments as well as on the DNA of complex genomes. The data in the first approach can be placed in a tridimensional matrix:

DNA fragment (F) x nucleotide element (E) x position (P)

In the second approach, the matrix is twodimensional:

F x Ep

where Ep are newly defined elements (i.e. oligonucleotide sequences) which integrate information about basic elements and their positions. Practically this means that the determination of a pisition is replaced bu the introduction of a substantially larger number of Ep elements instead of the basic elements E.

We propose to call the first approach EXPERIMENTAL and the second INFORMATIONAL. Since neither is purely experimental nor informational, the rationale for the choice is to name them by their dominating components. The INFORMATIONAL approach for obtaining and monitoring the primary structure of genomes can be defined as the basic principle of the methods which determine the presence of one, part or all of the oligonucleotides contained in specific genomic DNA fragment. The INFORMATIONAL approach can be used to determine the entire genomic sequence, or selected parts of a genome called GENOGRAM. Currently, the methods of this approach employ »mismatch free« oligonucleotide hybridization to experimentally determine the contents of oligonucleotides in genomic DNA. These methods are: polymorphic site detection[2,3,4], link-up method of obtaining ordered libraries of genomic fragments based on oligo hybridization[11,12] and the SBH method for determining genomic sequence[17]. Of course, the methods of the alternative EXPERIMENTAL approach exist for providing identical information. For example, RFLP analysis for polymorphic site monitoring[1,20], restriction site mapping and ordered libraries formation on the basis of overlapping restriction patterns[8,9,10], and gel electrophoresis techniques for DNA sequencing[13,14].

On the basis of the two informational possibilities, the methods of DNA sequence determination can be divided into those that determine which nucleotide is on which position in the linear sequence and the one which determine the contents of the olegonucleotide sequences necessary for the unique regeneration of a long sequence. Combinations of the two approaches are possible as well as the determination of the two positions of oligo sequences in the first and the determination of the contents of elements in the second. Both approaches can have variants in terms of acquiring unit data. Thus, the first approch is used ina gel sequencing (position is determined from the lenght of the DNA fragments)[13,14], sequentional removal and recognition of terminal nucleotides[6,7,15] and theoretically possible direct determination of nucleotides at every position by tunneling electron microscopy[6,7,16]. The second principle isused in SBH[17] in which the contents of oligo sequences in DNA are determined by oligonucleotide hybridization. One can pose the question whether the described difference in informational data matrices can serve as the basis for estimating which of the two approaches is more efficient and acceptable for sequencing complex genomes, or whether this estimate must always be based on technological considerations. Taking into account the presumed efficiency of IMAGE ANALYSIS, the question can be rephrased in terms of the estimate of the efficiency for obtaining the sum total of experimental images required in either approach. The unequivocal answer to this question is not yet available. In remainder of this paper, we develop an analysis of the INFORMATIONAL approach as groundwork towards this answer.

The INFORMATIONAL approach has several important methodological characteristics.

- The most basic one, as mentioned, is the possibility of acquiring experimental data bits about the contents of the oligonucleotides in unpositioned (of not necessarily determined position) samples. Since the sequence of elements is not the object of the measurements, nor is the physical distance of elements, there is no requirements for ordered spatial disposition of samples. For example, samples need not have the identical starting position as in gel electrophoresis.

- Second is that the contents of oligonucleotides can in principle be determined in samples which can have experimental micro volume or area. Since no physical separation of polymer molecules is required, transport effects such as diffusion, convection, etc. are avoided. These affects usually impose the necessity for macro volume or area.

- The third follows from the preceeding ones in that a considerable density of data bits is allowed per unit of experimental volume or area. Another way of phrasing this is the possibility of achieving a high degree of the parallel acquisition of experimental data.

- Fourth is the increase in the number of experimental data bits, i.e. increase in the size of the data matrix at the expense of the work needed to experimentally obtain a more sophisticated and maximally compact matrix. This means that the burden of overall efficiency is placed on the data reading step which has to input more data bits in computer

memory unit time. These characteristics permit the formulation of a concept for a miniaturized, fast and frugal process for the generation of experimental data sufficient for the determination and monitoring of primary structures of complex genomes.

B. THE CONCEPT OF WORKABLE METHODOLOGY

One can define four phases in the experimental part of current methods sharing the INFORMATIONAL approach, when the content of oligonucleotides in the DNA is determined by oligonucleotide hybridization. This part is followed by the informantional phase in which the computer generates a part or the entire sequence from experimental data. These four phases are:

1. Sample preparation (isolation, marking, preparation for phase 3)
2. Probe preparation (oligonucleotide synthesis, oligonucleotide bank, probe labeling)
3. hybridization (using one or two groups of differently labeled oligonucleotide)
4. reading the experimental data and storing them in the computer as crude information.

We have defined three goals which should be achieved either in each phase separately or in the process as a whole:

i) Samples and probes should be to the least possible extent recognizable by position (coordinates, addresses) and to the maximal extent possible by their innate informational characteristics.

ii) There should be the fewest possible number of independent hybridization reactions.

iii) The process should have a speed of delivering at least one million of +/- F(target) x Ep(probe) data bits per second to the IMAGE ANALYSIS.

B1. *Samples as a system of discrete and recognizable particles*

The samples are of importance in two phases: sàmple preparation and hybridization. Obvious solutions are the preparation of addressed samples in microtiter plates and addressed hybridization dots in ordered dot-blots. One can pose the question of whether the use of addressed samples can be avoided. This is especially important when the samples need not be kept for permanent storage, or can easily be obtained when needed so that it is easier not to keep them. The task is difficult when the detection system requires more than one target molecule, i.e.amplification of genomic fragments. Also, one cannot determine (or it is convenient not to determine to retain parallelism) the complete contents of oligonucleotides on a sample present only once in the hybridization area. Hence, it is useful to have a given genomic DNA fragment in a large number of copies in one hybridization spot and many such hybridization spots in separate hybridization areas. One would like to recognize such hybridization spots, even if they are not separated as amplified genomic fragments in tubes, with known marks and ordered hybridization dots with known coordinates.

As in the chemical synthesis of informational polymers, the answer is in the use of solid support as the substitute for reaction tubes to keep the soluble components apart. A drop of water solution is replaced by the solid particle (i.e. discrete particle (DP)) which is carrying the required number of copies of the same DNA fragment attached to its surface. These particles can be looked upon as small beads of defined size (and shape) similar to those already in use for DNA in different applications[21]. To simplify description, the use of DPs in a hybridization reaction will be presented first, starting from the physically separated and amplified samples obtained in phase one.

Starting from the genomic library consisting of clones already placed in microtiter wells, one adds DPs to each well and a binding reaction takes place, resulting in cloned DNA being attached to DP. Thus, each liquid sample is divided into a certain number of DPs. Aliquotes of DPs from each well can be mixed together and spread in a monolayer of required density. This is followed by fixation of the monolayer to a support. This way one can obtain hybridization areas (HA) similar to the filters in dot blot procedure. Every DP represents one dot, and a solid support a certain area of the filter. One can imagine a simple case in which each HA contains enough randomly displayed DPs that each clone is represented at least one. The other HAs have the same representation of clones but the DPs are displayed differently. In other words, in these »replicas«, the DPs are present in different places on the monolayer. Every HA can be hybridized and reused as standard filters. The main problem is to recognize a DP with the same clone in a different HA. This problem does not arise if all the hybridizations can be performed on a single HA. The latter is probably very difficult to achieve with the 10 oligo probes required for SBH. Also, the possibility of performing many hybridization rections in parallel is lost if only one HA is used.

We see three principal ways of recognizing DPs:

1. Labeling with the physical attributes of DPs such as size, shape and color.
2. Labeling with different combinations of oligonucleotides which can be recognized as such by hybridization with appropriate probes. They should be attached to DPs in addition to genomic DNA. For instance, one can label 10^7 DPs using a total of 26 oligonucleotides and forming combinations of 13 oligonucleotides. Thus, out of total number of necesary probes for hybridization with a given HA, 26

will always be used to recognize those DPs which carry the same clone.

3. Recognition of the DPs carrying the same DNA fragments by the use of a certain fraction of possible oligonucleotide probes in parallel hybridizations on all HA. The part of sequencing information obtained from genomic DNA on DPs is thus used for recognition purposes. This principle is not an obvious one and it is used in Lehrach's link-up method[11,12]. Lehrach found that 100 oligo probes having the same density as 8- or 9- mer probes for SBH are suitable for recognizing overlapped cosmids in an entire genomic library[11]. In the case here, one requires the recognition of identical and nonoverlapped clones in a mixture of a given complexity. The simpler the mixture, the smaller number of different probes it requires for recognition. The reguirement is that the HAs can withstand testing with a large number of probes, so that a part can be spared for recognition. All clones need not be mixed at once, but separate mixes with smaller numbers of clones can be prepared and used to obtain subdivided HAs, thus reducing the number of required probes.

With the combination of the three principles and the use of subdivided HAs where each division corresponds to simpler mixture, one can recognize the 10 separate samples required by GENOGRAM and SBH. It is obvious that the use of an HA divided into 10 parts and 100 »labels« per each principle allows the discrimination of this number of samples. In this combinatorial scheme, one has to prepare 10,000 samples with differently labeled DPs using principles 1 and 2. However, the maximal use of the third principle decreases the need to prepare differently marked DPs, but increases the number of required hybridizations.

1.B1.*Direct preparation of DP-genomic DNA complexes*

One can define three possible ways of making samples as direct mixes of DPs. The choice depends on whether the detection of oligonucleotide contents is done on one or a large number of DNA fragment molecules and on the mode of amplification. Their use ensures the elimination of macro-separated and addressed samples.

1. The attachment of an individual DNA molecule to each DP. Irrespective of the mode of attachment, the following informational property is important. There are multiple copies of each sequence present on various fragments in a mixed binding reaction. Therefore, DPs recognizable by their intrinsic properties can and will bind the same sequence fragment's. The way to offset the inability to use DP labeling principles 1 and 2 is to bind to DPs distinct by these labels physically separated parts of genomic or individual DNA (i.e. chromosomes, chromosome fragments, YACs, atc.) in separate reactions. The DPs obtained are mixed afterwards and used to form one HA. DPs carrying the same fragments will be recognized using principle 3 from the previous section. This is used only on DPs that appear the same according to principles 1 and 2 and come from a defined part of a genome or individual DNA. In contrast to cloned DNA, the fragments carrying the same sequence in this application are rarely identical, so that groups of densely overlapped fragments are recognized. The complete contents of oligonucleotides are obtained only for shared sequences.

However, at least in part, previously mixed cloning and/or amplification to supply DNA for DP binding is still necessary for some applications. For instance, in order to use random groups of fragments obtained by ligation (ordering library in SBH[17]), one needs to PCR or clone them. The »separation« of DNA required for a GENOGRAM from the rest of genomic DNA is best accomplished by a PCR reaction.

2. DP preparation with *in vivo* and *in vitro* amplified DNA. Of the many possible ways to do this, the alternative of obtaining genomic libraries using PCR is especially interesting. The maximal length of its insert is determined by the success of amplification. Lenghts of 6-10 kb have been reportad[22,23]. The procedure would require the ligation of a single primer to the ends of the genomic fragment mixture, the dilution of the ligation products to single molecules per desired volume, and then their use in separate PCR reactions, e.g., in microtiter wells. In this way, clones of starting fragments could be obtained *in vitro*, thus eliminating the uncontrollable aspects of *in vivo* cloning.

It is possible to see the implementation of PCR without the separation of individual fragments in distinct reaction volumes (see APP. 1).

3. The saparation of groups of densely overlapping DNA fragments on DPs capable of selecting DNA, instead of amplification. One can imagine separation by selection on the basis of a hybridization similar to the use of poly dT columns for the selection of mRNAs containing poly A. One would need 10 million DPs having different oligonucleotides satisfying the requirement of being represented only at once in the genome. An example for this is given in APP. 2 and APP. 3.

Procedure 1. is the most simple in a technological sense, but the detection of hybridization on a single molecule is a difficult, still unresolved problem. One possible solution is outlined in APP. 4. The other disadvantage is that recognition of identical fragments arising from the same physically separated part of the genome can only be done using principle 3. This in turn requires a large number of different hybridizations per single HA. The other

two procedures involve many technically untested operations. On the other hand, the conceptual definition of three separate but possible procedures allows some optimism that the mixed reaction for obtaining defined genomic DNA fragments on recognizable DPs is a goal that can be achieved.

B2. *Oligonucleotide bank: synthesis in mixed reactions*

The synthesis of a large number of separate oligonucleotides required as probes by the informational approach is a considerable task if standard »gene machines« are used. However, this synthesis can be appreciably speeded up using combinatorial principles[21]. This approach ensures the more rational and cheaper synthesis of smaller quantities of individual oligonucleotide probes. An even higher degree of rationalization can be achieved by the synthesis of sufficent quantities of large numbers of different oligos having multiple and repetitive applications which could be used by different laboratories (APP. 5). This is the concept of the oligonucleotide bank (an initiative by Crkvenjakov, Drmanac and Beattie). One can ask the question which bank would be the most useful.The answer lies in the recognition of oligonucleotide characteristics that are the most suitable for major areas of their possible uses. These areas are the detection of sequence by hibridization, the directed change of existing DNA molecules and the synthesis of a DNA of designed sequence.

The detection of sequence on short separated DNA fragments (amplified fragments, subclones, clones suitable for SBH) can be performed even with very short oligos, 8, 7, or even 6 nucleotides in length[24,25], our unpublished data). Probes about 20 bases long are suitable for hybridization with genomic DNA[26,27]. Primers for site specific mutagenesis and PCR are usually 15-20-mers[3]. Even 8-mers are active primers in PCR (Wallace, private communication). The procedure for DNA synthesis based on the sequential joining of short blocks is being developed[21]. We consider a bank containing all possible 3-8- mers very useful for the following reasons:

i) cited areas of application,

ii) technologically acceptable number of samples for the bank to contain, since there are about 90,000 different 3-8-mers.

iii) the possibility of generating longer sequence from shorter blocks present in the bank (ligation, the use of dideoxynucleotides and terminal transferase). Also, the advantage of making a bank on a solid support has been noted by Beattie. These oligos would lend themselves easily to further modification or extension, and if the solid support beads are suitable as DPs, no detachment procedure need be used. In the use of mixes of differently labeled probes as in SBH or other applications, one can synthesise mixes and not form them by mixing. For example, all 8- mers can be synthesised in 1024 reactions, if there are 64 different nucleotide labels. All 64 3-mers are synthesised in large quantities on DPs and are differently labeled. One thousand and twenty-four of equimolar parts of the 64 samples are made and a specific pentanucleotide is synthesised on each of them.

A similar principle could be used for obtaining 10^7 DPs, where each is carrying different longer oligonucleotides (APP. 3). These are necessary for the preparation of sample mixtures for selection by hybridization (see section B 1.1.). The savings resulting from the oligonucleotide bank and the use of DPs are outlined in APP.5.

B3. *Image analisys: The use of group of differently labeled probes*

As mentioned earlier, the INFORMATIONAL approach of moderate efficiency needs to store 10^5 to 10^6 bits of F, E_p binary information per second in a computer memory. Since the matrix for this approach consists of 10^7 targets x 10^5 probes, this speed can be attained at the two extremes: by reading 1 target with all the probes per second, or 10^5 to 10^6 targets with a single probe per second. Of course, all matrices within the two extremes are possible as well. On one end, it is necessary to perform 100,000 separate hybridizations; on the other, simultaneous hybridization followed by the recognition of 100,000 different probes. Therefore, the parallel formation of many submatrices of the type 10^4 to 10^5 targets x 10-100 probes seems to us to be the most rational way to proceed. The parallelism can be of the two kinds: a formation of all the required 100-1000 submatrices on one HA and scoring with single probe group, and/or a parallel hybridization in separate vessels on HA »replicas« with many different probe groups. In either case, it would be more desirable to use groups of 100 differently labeled probes which would require only 1000 separate hybridizations and as many separate oligonucleotide synthesis reactions (see section on the oligo bank). On the other hand, this would require the use of 10^4 to 10^5 tergets per experimental image. Electronic cameras would have to contain about 10^6 pixels to achive efficient image analysis of that many targets. CCD cameras can have from 650,000 to 1.3 million pixels of about 10x10 μm and therefore, can fulfill this requirement[19]. The use of probe groups and some problems in labeling are treated in APP.6.

DISCUSSION

This work represents an attempt to define a more rational approach to the development of the methods necessary in the resolution of central problems, in molecular biology such as the determination of primary genome structure, etc. Theoretical treatment of the problem and the comparison of the properties of all possible approaches has the goal of discarding some methods and procedures as inefficient or impossible and initiating and facilitating the development of those that show

theoretical promise. One can pose the question of why the PCR reaction has not been developed earlier when all its material components have been known for over a decade. It is tempting to assume in retrospect that a thorough theoretical analysis of the ways of amplifying single DNA fragments or fragment libraries could have predicted PCR with its now plainly obvious advantages. Would the existence of such a theoretical concept have led to the earlier emergence of PCR? Due to the complexity and size of the hyman genome project, it inevitably needs theoretical treatment of the methodological requirements and the ways of meeting them.

The INFORMATIONAL approach defined here is based on the use of oligonucleotide »words«. Interestingly, the same lexic principles form the basis of the most efficient algorithms for sequence comparisons existing today[28,29]. For the moment, there are two experimentally realistic ways to determine the contents of oligonucleotide »words«. These are based on a naturally occurring molecular processes: base pairing of nucleic acids and sequence specific protein-nucleic acid binding. The base pairing is more general, since there is no restriction on the specific sequences involved and probably is easier due to hybrid stability. In light of this, we believe that oligonucleotide hybridization will have central role in the compilation of genomic sequences.

The fundamental principle of the INFORMATIONAL approach provides the technological advantages of unbroken parallelism and amplification cascades. Unbroken parallelism means that samples are not separated from each other from genomic DNA to image analysis stages. Amplification cascade is represented by 10 genomic parts x 10^{5-6} DPs x 100 HA x 10 washes x 100 probes/hybridization x 100 hybridization/day x 10 days yielding 10^{13} units of information data in 10 days. We would like to call the proposed parallelism quadratic parallelism. A small number of operations are performed in each step, but due to the multiplication and not the summation of gains of unit unformation bits per step, the total yield of information bits is enormous.

The INFORMATIONAL approach also has an additional technological advantage of a general nature. It permits the maximization of the use of resources and materials prepared beforehand that are identical for different samples and the nininization of sample specific treatments. These functional cassete preparations are independent of the source of genomic DNA and usable for the compilation of sequence information from any individual or species. The three cassetes are:

i) The probe bank containing the integrated information of type-basic element x position, usable for any DNA allowing the corresponding decrease in experimental determination of information.

ii) The DP preparation integrating the addressing information, so that it need not be determined by robotic positioning operations.

iii) The computer software package for the generation of sequence from the required input of experimental data, which once made is reusable on new data sets without the intervention of the experimentalists.

In the final analysis, the INFORMATIONAL approach provides a decrease in experimental requirements for the genomic sequence compilation at the expense of informational computer work. The following is a tentative estimate of savings in implementing some of the concepts introduced here for an efficient INFORMATIONAL approach. The experimental surface are as well as the sample and probe mass requirements are decreased 1000-fold by the proposed miniaturization (APP. 5). Even more important that space and material savings are reductions in the number of robotic operations. A robotic hand with 10000 pipetting units needs 1000 operations to make one filter with 10 million dots. Using the DP system, a robotic hand with several pipetting units can perform an analogous task in a single operation. The practical realization of miniaturization concepts can lead to genome sequencing equipment of the size of large present day instruments for a molecular biology laboratory.

The DP system in essence represents the imitation of a multitude of biochemical reactions occurring simultaneously inside cells. The specificity and discreteness of cellular reactions are based on enzyme actions whose informational properties are imitated here by DPs. The use of DPs requires an icrease of an order of magnitude in the number of unit information bits, but time and labour investments for obtaining the complete data set are reduced several-fold in respect to other procedures. In a robotized dot system, every DNA fragment is represented in each filter only once. This certainty is replaced with the probability that each clone is represented at least once in the HA. This imposes a 10-fold increase in the number of DPs over addressed samples. On the other hand, this increase allows the tolerance of imperfect hybridization on individual DPs, i.e. statical determination of positive hybridization. This last instance means the reduction of required experimental performance level. Therefore, the DP based hybridization and signal reading procedures must tolerate the libraries consisting of a larger number of fragments. This in turn allows the use of a smaller number of probes, which is especially evident in GENOGRAM application. Instead of specically choosing 10,000 pairs of primers and probes, the DP system allows the use of all amplified fragments with all probes at once. The advantage is further strengthened by the realization that the ensuing surplus of information means a higher accuracy in polymorphism determination as well as the possibility of the detection of new mutations, both of which can be of considerable diagnostic value. The switching of emphasis from experimental work to the image analysis in the DP approach allows the determination of a greater number of detailed GENOGRAMS than in previous procedures.

The properties of the INFORMATIONAL approach described here, besides miniaturization, provide for a greater time savings than other methods which retain the requirement of the experimental determination of more complex and variable types of data.

It is interesting to attempt to outline in comparative analysis the advantages and disadvantages of the INFORMATIONAL approach for genomic sequencing versus three main procedures which use the EXPERIMENTAL approach. These are position determining methods. The standard method used up to now, based on the finding of the position by measuring the lenght of DNA fragments, has two requirements which almost certainly exclude it as a method of choice. These are the practical impossibility of miniaturazation and the need for the use of amplified DNA fragments. The other two methods which, like SBH or other procedures using the INFORMATIONAL approach, have not been experimentally verified do not impose these requirements. The tunneling electron microscopy is an inherently miniaturized procedure which does not require amplification. Also, the sequential removal of a nucleotide by a nucleotide from one end of a DNA fragment, followed by continual separation by flow and efficent registration in passage by the detector, for practical reasons almost certainly requires a single DNA molecule. It is very difficult to synchronize the removal of nucleotides on the level of fentomoles of identical fragments. It can be imagined that this procedure can be miniaturized and made parallel, since instead of the addressed samples, multiple microtubes can be used ensuring discreteness. The separation of nucleotides from DNA by principles akin to those applied in flow cytometry does not require macro separation similar in precision to gel electrophoresis. The main requirement of this approach is precision and speed of detection of single events, especially parallel detection in a large number of microtubes. The question is can the use of lasers and fluorescence labeling combined with pixel-based image analysis allow an acceptable data acquisition speed with nonprohibitively complex equipment. In any case, both discussed procedures of the EXPERIMENTAL approach rely on the development of sophisticated physics equipment while the INFORMATIONAL approach is exclusively based on biomolecular processes. Since there is an indirect detection of molecular reactions, the SBH does not need to »see« atoms and since it does not use position information, the SBH does not require any physical ordering of rections, allowing the use of amplified fragments. Therefore, the technical requirements for reaching the image analysis stage in SBH and other procedures based on an INFORMATIONAL approach are minimised in comparison with the alternative EXPERIMENTAL methods, as shown here in detail.

All procedures for genome sequence compilation have a common last step, CCD camera analysis of experimental images. Thus, comparative judgment has to be based on the criterion of the case of obtaining the total experimental images for sequence determination. It appears to us that SBH is more adapted for sequencing a large number of complex genomes than the other methods discussed here. Large-scale sequencing has scale- dependent requirements. Through the use of aliquotes of once prepared cassetes, SBH approaches the ideal technological requirements for an efficient method. This reduction of individual genomes to a common denominator, oligonucleotides, allows the use of informational work after image analysis for sequence generation. In the EXPERIMENTAL approach based methods, most of the work is experimental in nature.

The purpose of this work is to enumerate some specific properties and requirements for procedures based on an INFORMATIONAL approach and to argue the case for its theoretical advantages over the alternative approach. There is a need for further work on both the theoretical and practical possibilities of all the methods mentioned before overall advantages can be determined. It is quite possible that both approaches have advantages in specific domains of the genome project, so that finally the most efficient method might be based on their complementarity.

APPENDIX

APP.1. *On the possibility of mixed amplification by PCR*

The requirement is that micro droplets, each containing either a single apmlifiable DNA fragment or none, are enclosed into semipermeable spheres together with DP aggregates and the necessary PCR components. These aggregates should be separable into individual DPs under mild conditions. The use of aggregates provides a way to prepare more DPs with the same DNA fragments. Multiple DPs are required for multiple HA »replicas«. Microsphere formation should be considered as a statistical process with a certain degree of success rather then a highly robotized process with high fidelity. Every microsphere represents a separate amplification reaction similar to a microtiter well. The reaction of the binding of amplified fragments to DP aggregates would require the use of a reagent for which the microsphere is permeable. Following binding, the disruption of the microsphere membranes and DP aggregates would take place in a common reaction. This would result in a mix of DPs in which each DNA fragment is represented in a sufficient number of copies on an adequate number of DPs (»cloning«).

APP.2. *On the possibility of »cloning« by hybridization selection*

It is necessary to have 10 million DPs, each carrying a specific oligonucleotide. The later will have lengths which ensure theire occurrence mostly once in a genomic sequence. The ways of obtaining this number of different DPs in parallel are treated in B2 and A3. The genomic DNA fragments

suitable for cloning by hybridization selection are obtained in the following way. The large mass of random DNA fragments obtained by shearing and size selection to twice the lenght finally required are exposed to the limited action of 5' or 3' exonucleases. The lengths of single - stranded tails thus obtained should be in the range of 100-1000 bp. These fragments are subsequently randomly cut and again size selected. The resulting fragments of required unit size have a single stranded tails on one end only. They are hybridized to DPs. After hybridization, the hybrid is filled in and ligated. In this way, each DP will have bound to itself only those fragments which are internally displaces maximally for the length of a single-stranded region containing the DP specific oligonucleoticd. The recognition of DPs with the »same« fragments can be done by labeling using any one or a combination of principles treated in B1. One would need 0.1-1 g genomic DNA for this application. This selective procedure is even more applicable to a GENOGRAM where the number of samples per indiviudal genome is 1000-fold smaller than in SBH. DPs would carry specific sequences complementary to the ones that need to be examined in GENOGRAM analysis.

APP.3. *The preparation of probes and their use in hybridization selection*

We think that it is sufficient to have 1000 different DPs recognisable by principles 1 and 2 (see B 2). On each DP, a different 5-mer is made in 1000 separate syntheses on a DNA synthesizer. About 100 equimolar mixes of these 1000 are used for further 5-mer extension. This process is repeated once again for another 6 nucleotides. In this way, only 1200 syntheses are necessary to obtain the 10 different DPs necessary for the formation of the same number of »clones« by selection. The use of such DPs allows the sorting of the obtained »clones« into 1000 groups. The sorting principle is based on DP characteristics (physical properties; principle 1 and oligonucleotide combinations; principle 2). The same »clones« always belong to a single group. The detection of identical »clones« would thus be restricted to a group and would be accomplished by comparing the oligonucleotide contents obtained with a subset of the probes used in all HAs for identification purposes (principle 3). This should not be a problem, since each group contains 10 »clones« and there is considerable chance that a majority of overlapped »clones« groups.

APP.4. *On the possibility of oligonucleotide detection in single DNA molecules*

If one considers hybridization alone as a means for the detection of oligonucleotides in a single target molecule, the problem has two components. The first is the probability of the occurrence of the hybridization event with a single target with excess probe. The second is the detection of the obtained hybrid. Since no efficient or simple procedure for the detection of single molecule hybridization has been developed so far, there is no knowledge of this reaction either.

If one assumes that the event of single molecule hybridization occurs with a certain probability, the detection can be of two kinds. In the first, the detection of the signal is produced by the marker on the hybridizing probe (e.g. fluorescence, enzymatic activity, even if later amplified in various way[15,30]). However, as mentioned, the detection of a single hybridized oligonucleotide molecule has not been reported yet. In the second kind, the hybridization event is amplified itself. Its logic is the same used in all exponential doublings in natural and *in vitro* amplification reactions (cell division, DNA replication, PCR[3], ligation-amplification reaction (LAR)[5]). The final yield of product is $k \times 2^c$, where c is the number of cycles and k the efficiency factor. For GENOGRAM determination, one can use LAR, since the main requirement of the method is obeyed which is the previous knowledge of the sequence or the small number of its variants. LAR is more difficult to apply to an SBH of unknown sequence. An oligonucleotide, the reporter of a hybridization event (carries biotin in cited applications), would have to be very short in order to use all the theoretically possible sequences in a mixture, i.e. probably 4 to 5 mer. In addition, LAR has the problem of how to localize the ligation product on the dots or DPs on which it is formed. If the problem of local fixation is resolved, one can avoid the requirements of the specificity of the ligation reaction and reporter molecules in the following manner.

DPs carrying the capability of binding oligos having a specific chemical group and a single target DNA fragment are prepared. Then one hybridises with a probe that is both complementary and carries the chemical group. After hybridization and washing, the reaction of denaturation and binding is performed. It is necessary to permit the binding of oligonucleotide hybridized to the target on a given DP only. In this way, a DP having a positive hybridization would present two targets in the next round of reactions. One repeats these cycles until a detectable number of oligos is bound to DPs. It is interesting to note that after the first cycle which should be with short oligos (8-mers for SBH), one can switch to hybridization with longer probes and/or targets quite independent form the first target sequence. The discrimination is thus easily achieved in a slightly higher number of cycles. This is accomplished by the use of synthesized targets which are longer than a primary oligonucleotide and by the use of additional oligonucleotides complementary to a synthesised target.

This scheme is just a theoretical possibility in the detection of single molecule hybridization and does not presume experimental feasibility. In any case, the detection of hybridization on the level of a single molecule, being a process with a small probability of positive outcome, can be treated statistically as a result of trials on a larger number of DPs with the same target or a larger number of trials with same oligonucleotide on the same DP.

This larger number of trials is integrated in repeated cycles of hybridization. The positive hybridization is recognised within a wide span of signal intensities above a certain threshold. This is similar to the situation when in various DPs or dots there is a large difference in the molarity of amplified targets.

APP.5.*Savings in using an oligo bank and the process based on DPs*

According to Beattie's calculation, a bank of 8-mers (65,536 oligos) could be synthesised in less than 6 months with atotal investment of 3 million dollars. Since the cost of materials and labor for a such bank, having a stock of 2 mg of each oligonucleotide, is 2 million, 10 g quantities of all probes sufficent for an efficient SBH would cost 10-20,000 dollars, some 1000-fold less than the present commercial price. If the full possibilities of an oligonucleotide bank and combinatorial synthesis are used, one can see that the cost of the synthesis of oligonucleotides could be directly proportional and only slightly higher than the cost of material, i.e. nucleotides. The use of DPs instead of filters also has great advantages in terms of cost, since their use requires a smaller quantity of oligo probes.

For the same target density per unit area, there is a much smaller quantity of target genomic DNA per DP than per dot. This comes from the fact that smaller discrete areas are imaged in DP than in dot-based hybridization. There is also a decreased use of buffer and probes since the total HA area is much smaller. The area of the largest section of DP is $10\mu m^2$, assuming a DP diameter of $4\mu m^2$. If a dot area is 1 mm^2 , then the ratio of areas is $1:10^5$. For the formation of a monolayer, one can assume the use of a 10-fold higher number of DPs is necessary to permit each DP to appear at least once in the HA, and that the use of space by productive DPs is 10%, with a ratio of 1:1000. For example, the HA area in the DP-based process is 10 x 10 cm versus 1 x 1 m in the dot blot one. Assuming that one HA can be scored by 1000 probes (a mixture of 10-100 probes and 10-100 washes) the total area is 1 m^2 versus 1000 m^2. The required quantities of probes could be calculated in this way. One probe has a target in every tenth clone. This factor is cancelled by another factor of 10 due to a random sampling. Therefore, the total number of DPs with which one probe hybridizes is equal to the total number of clones, i.e. 10. At best, for the detection of a signal from one DP, one would need no more than 1000 fluorescing molecules. Even assuming that only 0.1% of the probe molecules are taken up into the hybrid, only 10 molecules will be necessary for hybridization with all the clones. Since 1 μg of 8-mer contains about 3×10^{14} molecules, this quantity should be sufficient. The dot blot system would require a mass of probe of the order of 1 mg.

The savings per one genome sequences, or 100 GENOGRAMS determined, using the oligonucleotide bank cost estimates of Baettie, would be on the order of 1 million dollars in switching from the dot blot to the DP-based determination of oligonucleotide contents.

Similar savings are obtained for the volume of cloned samples. In macro applications based on existing technologies of gel sequencing or the dot blot system, at least 100 μl per clone are required. For 10 million clones, this amounts to one ton of liquid. For the microhybridization, concept this volume is reduced to several liters at most.

APP.6.*On the possibility of using multiple-labeled probes simultaneously in hybridization*

In the most suitable matrix, there is a formidable problem. In simultaneous hybridization with 10-100 probes, one needs that many labels recognizable in the experimental image. One should also keep in mind that the theoretical expectation is that each probe will hybridize with 10% of the targets, or even less in the GENOGRAM. This means that with 100 probes, 10 different labels will be present per each DP or dot. There are two methodologies devised for similar problems which can be used for the INFORMATIONAL approach as well. One is the use of different fluorescent dyes[31,32], and the other is gas chromatography coupled with mass spectroscopy[33]. In the former, it is difficult to imagine the use of more than 10 fluorescent dyes at once and also for their detection on a single target one needs an exciting light of different wavelengths and/or different filters, both of which can slow down the image collection process. Every change of wavelength or filters is an additional physical operation, making optimization possible only in hybridization and not in image analysis. However, the extreme precision and sensitivity of CCD cameras (down to two photons per second) can overcome this disadvantage. The gas chromatography-mass spectroscopy approach can potentially discriminate even 1000 labels, but is possible only on single or a limited number of samples at once. For this methodology to work, one would need to develop a technology of the parallel acquisition of total data from 100-1000 samples per second. This is difficult to blend with use of unordered microsamples in the form of DP.

The most simple for image analysis is to discriminate between objects on the basis of their size, shape and color. This reduces to different photon patterns in contrast to defined photons emitted by fluorescing molecules and permits the recognition of an »unlimited« number of nonoverlapping objects. Therefore, the recognition of 100 physically different DPs in image analysis must be simple. One can ask the question whether this principle of target labeling can be used for probe labeling as well. Two principles can be applied: i) the probe carries a physical entity recognisable by optical microscope and thus usable in image analysis and ii) the probe carries a chemical entity which can be used after hybridization for the localized formation of a specific physical entity.

The most simple application of the first principle is the double DP system where probes as well as targets are bound to DP which can be separately

recognized. Positive hybridization would lead to rosette formation. The target DP would be surrounded with DPs carrying probes whose complementary sequences are present in a given target. This principle of the visualization of a bimolecular recognition reaction is seccessfully employed in antigen-antibody reactions. Cell-cell rosettes are well known in immunology, while microbead rosettes have been used occasionally. The practical question is can oligonucleotide hybridization lead to the formation of a sufficicent number of weak chemical bonds whose total energies are strong enough to hold two DPs together. In this context, one should not forget that the applied system must allow hybridization discrimination on the level of one base mismatch.

The alternative approach does not require the linking of DPs by the chemical bonds formed in hybridization. Its problem is how molecular initiating event can be transformed into a locally recognisable physical character. One should probably take advantage of a local concentration of certain reagents (say a certain metal ion). One can rephrase the question in the following way: how can one transform in one or more reactions the 10 different localized entities recognizable by optical microscope. One can speculate on the ionic initiation of chemical reactions on the surface of target DPs with labeling DPs added in the system after hybridization. The other possibility is the initiation of the formation of recognizable microcrystals.

We think that it is important to discuss all possible approaches to probe labeling, irrespective of their immediate applicability in experiments, in order that development can be directed to the potentially most efficient method in both the theoretical and practical senses.

ACKNOWLEDGMENTS

This work has been supported in part by the Self-Management Community of Interest for Science of Serbia, DOE Grant DE-FG02- 88ER60699 and US-Yugoslav Joint Board Project JF 820. This work is the subject of Yugoslav patent applications.

REFERENCES

1. BOTSTEIN D WHITE R L SKOLNICK M DAVIS R W *AM J HUM GENET* 1980 **32** 314-331
2. WALLACE R B SHAFFER J MURPHY R F BONNER J HIROSE T ITAKURA K *NUCLEIC ACIDS RES* 1979 **6** 3543-3557
3. SAIKI R K BUGAWAN T L HORN G T MULLIS K B ERLICH H A *NATURE* 1986 **324** 163-166
4. LANDERGRON U KAISER R SAMDERS J HOOD L *SCIENCE* 1988 **241** 1077-1080
5. SKOLNICK M H WALLACE R B *GENOMICS* 1988 **2** 273-279
6. Congress of the United States Office of Technology Assessment Mapping Our Genes- *The Genome Projects : How Big How Fast?* OTA-BA-373 Washington D C U S Government Printing Office April 1988

   National Research Council *Report* of the Committee on Mapping ond Sequencing the Human Genome 1988 National Academy Press, Washington DC

   U S Department of Energy *Report* on the Human Genome Initiative for the Office of Health and Envirnomental Research Livermore CA 1988
7. SMITH L HOOD L *BIO/TECHNOLOGY* 1987 **3** 933-939
8. OLSON M V DUTCHIK J E GRAHAM M Y BRODEUR G M HELMS C FRANK M MACCOLLIN M SCHEINMAN R FRANK T *PROC NATL ACAD SCI* USA 1986 **83** 7826-7830
9. COULSON A SULSTON J BRENNER S KARN J *PROC NATL ACAD SCI* USA 1986 **83** 7831-7825
10. KOHARA Y AJUYAMA K ISONO K *CELL* 1987 **50** 495-508
11. POUSTKA A POHL T BARLOW D P ZEHETNER G CRAIG A MICHIELS F EHRICH E FRISCHAUF A-M LEHRACH H *COLD SPRING HARBOR SIMPOSIA ON QUONTITATIVE BIOLOGY* 1986 **51** 131-139
12. MICHIELIS F CRAIG A G ZEHETNER G SMITH G P LEHRACH H *CABIOS* 1987 **3** 203-210
13. SANGER F NICKLEN S COULSON A R *PROC NATL ACAD SCI* USA 1977 **74** 5463-5467
14. MAXAM A M GILBERT W *PROC. NATL ACAD SCI* USA 1977 **74** 560-564
15. NGUYEN D C KELLER R A JETT J H MARTIN J C *ANAL CHEM* 1987 **59** 2158-2161
16. BINNING G ROHRER H *SCI AM 1985-253 50*
17. DRMANAC R LABAT I BRUKNER I CRKVENJAKOV R *GENOMICS* 1989 **4** 114-128
18. BLOUKE M M JANESICK J R HALL J E COWENS M W MAY P J *OPT. ENG* 1983 **22** 607
19. HIRAOKA Y SEDAT J W AGARD D A *SCIENCE* 1987 **238** 36-41
20. DONIS-KELLER H et al *CELL* 1987 **51** 319-337
21. BEATTIE K L LOGDSDON N J ANDERSON R S ESPINOSA-LARA J M MALDONADO-RODRIGEZ R FROST J D III *APPL BIOCH* 1988 10
22. SAIKI R K GELFIND D H STOFFEL B SCHARF S J HIGUCHI R HORN G T MULLIS K B ERLICH H A *SCIENCE* 1988 **239** 487-491
23. JEFFREYS A J WILSON V NEUMANN R KEYTE J *NUCLEIC ACIDS RES.* 1988 **16** 10953-10971
24. ESTIVILL X WILLIAMSON R *NUCLEIC ACIDS RES* 1987 **15** 1415- 1425
25. LEE B L BLAKE K R MILLER P S *NUCLEIC ACIDS RES* 1987 **16** 10681-10697
26. CONNER B J et al *PROC NATL ACAD SCI* USA 1983 **80** 278-282
27. THEIN S L WALLACE B *Human Genetic Diseases: a Practical Approach* Ed Davies K E Oxford IRL Press Limited 1986 33- 50
28. LIPMAN D J PEARSON W R *SCIENCE* 1985 **227** 1435
29. KARLIN S MORRIS M GHANDOUR G LEUNG M-Y *PROC NATL ACAD SCI* USA 1988 **85** 841-845
30. URDEA M S WARNER B D RUNNING J A STEMPIEN M CLYNE J HORN T *NUCLEIC ACIDS RES* 1988 **16** 4937-4956
31. SMITH L M SANDERS J Z KAISER R J HUGHES P DODD C CONNELL C R HEINER C KENT S B H HOOD L E *NATURE* 1986 **321** 674-679
32. PROBER J M TRAINOR G L DAM R J HOBBS F W ROBERTSON C W ZAGURSKY R J COCUZZA A J JENSEN M A BAUMEISTER K *SCIENCE* 1987 **238** 336-341
33. ADAMS J DAVID M GIESE R W *ANAL CHEM* 1986 **58** 345-348