Published online February 18, 2004

1208–1213 *Nucleic Acids Research, 2004, Vol. 32, No. 3*
DOI: 10.1093/nar/gkh210

A gene-specific DNA sequencing chip for exploring molecular evolutionary change

Olivier Fedrigo and Gavin Naylor[1],*

Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011, USA and [1]School of Computational Science and Information Technology (CSIT), Florida State University, Tallahassee, FL 32306-4120, USA

Received September 9, 2003; Revised November 29, 2003; Accepted December 15, 2003

ABSTRACT

Sequencing by hybridization (SBH) approaches to DNA sequencing face two conflicting constraints. First, in order to ensure that the target DNA binds reliably, the oligonucleotide probes that are attached to the chip array must be >15 bp in length. Secondly, the total number of possible 15 bp oligonucleotides is too large (>4^{15}) to fit on a chip with current technology. To circumvent the conflict between these two opposing constraints, we present a novel gene-specific DNA chip design. Our design is based on the idea that not all conceivable oligonucleotides need to be placed on a chip—only those that capture sequence combinations occurring in nature. Our approach uses a training set of aligned sequences that code for the gene in question. We compute the minimum number of oligonucleotides (generally 15–30 bp in length) that need to be placed on a DNA chip to capture the variation implied by the training set using a graph search algorithm. We tested the approach *in silico* using cytochrome-b sequences. Results indicate that on average, 98% of the sequence of an unknown target can be determined using the approach.

INTRODUCTION

The use of DNA sequencing as a tool to shed light on biological processes has increased dramatically in the past 15 years. Sequencing technology is currently used in a diversity of fields: from forensic science to phylogenetics, from conservation genetics of endangered species to tracking the evolution of viral epidemics. As we move into an era of comparative genomics, drug discovery and genetic screening, the demand for high throughput approaches to DNA sequencing is becoming ever more pressing (1).

Because traditional sequencing methods based on gel electrophoresis are both costly and time consuming, several alternative approaches are being explored. One of the most promising is sequencing by hybridization (SBH) (2–4). This approach, like PCR, takes advantage of the fact that DNA binds to its complementary sequence. Briefly, a 'chip' is made by immobilizing several thousand short pieces of DNA of known sequence (oligonucleotides) to a glass or silicon surface at pre-specified locations. The chip is then incubated with fluorescently labeled target DNA whose sequence is to be determined. The sequence of the target is deduced from the subset of oligonucleotide probes that are bound, which can be identified by their change in color (fluorescence). This information is then passed to a computer and the target sequence is reconstituted from the patterns of oligonucleotide bound to the target (5). The chemistry of such hybridization reactions generally requires that oligonucleotides of length 15–30 bp be used to ensure reliable binding (4,6,7). However, the total number of combinatorial variants for an oligonucleotide of length 15 (4^{15} or 1 073 741 824) is too large to fit on a single DNA chip, which is currently constrained to hold about 400 000 oligonucleotides. The conflicting constraints imposed by chemistry and the limited number of oligonucleotide variants that can be housed on a chip have prevented development of universal sequencing chips. Some improvements have been proposed. At the chemistry level, the use of semiconductor technology has been proposed to increase the capacity of the chip (8), while universal bases that bind to all four nucleotides have been proposed to decrease the number of probes required to capture all conceivable combinations (9). At the post-processing level, algorithms have been developed to improve the accuracy with which the target sequence can be reconstructed from bound probe information, which means that chips with shorter, and therefore fewer, oligonucleotide probes can be designed (10). Unfortunately, these improvements have not been sufficient to overcome the obstacles associated with development of a universal SBH sequencing chip. This is not to say that SBH does not have practical applications. It does, but in a much more restricted setting than would be the case if a universal sequencing chip could be designed. SBH approaches are currently used in screening for specific target sequences known to be associated with genetic diseases such as cystic fibrosis (11).

To recap, universal sequencing chip design requires that two conflicting demands be reconciled. Probes must be long enough to bind reliably, but not so long that the number of probes will no longer fit on a chip. In this study, we do not offer any solutions to the universal sequencing chip problem, but focus instead on a subproblem that lies somewhere along the continuum between the development of specialized chips

*To whom correspondence should be addressed. Tel: +1 850 645 0314; Fax: +1 850 644 0098; Email: naylor@csit.fsu.edu hardwired to detect pre-specified mutations and the universal sequencing chip design problem. Rather than trying to represent all possible oligonucleotides of a given length necessary for a universal sequencing chip, we set out to develop gene-specific chips tailored to the re-sequencing phase of genomics, for phylogenetics and population genetic screening, viral epidemiology and forensic applications. By restricting our focus to particular genes, we are able to reduce the number of oligonucleotides that would be required for a universal sequencing chip. Our goal is more narrowly defined than that of universal sequencing chip design but more flexible than the specialized chips used to detect particular sequences in genetic screening applications. The design problem for a gene-specific sequencing chip is: how do we estimate the minimum number of probes that will capture the variation that occurs in nature for a particular gene? The solution we advocate herein uses an approach that leverages information about natural variation contained in multiple alignments.

DEVELOPMENT OF A GENE-SPECIFIC CHIP

Combinatorial approach

Multiple sequence alignments reveal conserved and variable sequence regions. This distribution of sequence variability reflects structural and functional constraints. Regions that are tightly constrained are less free to vary than regions which are not as constrained, all else being equal. However, the relationship between functional constraint and variability is not absolute. Some regions can be functionally constrained, but still free to vary for a subset of amino acids that do not jeopardize function. For any given region, there is a finite set of amino acid combinations that can be substituted without changing structure or function. If we could estimate this for all regions of a protein, we could estimate all possible variants viable for a particular protein's function. This is the principle underlying our gene-specific sequencing chip design. Because *ab initio* prediction is not yet possible, we use training data sets based on sequences that occur in nature to estimate the sequences that would be viable for a particular protein. A training set comprises a multiple alignment of sequences taken from a diversity of organisms for a given protein-coding gene. The multiple alignment gives an immediate indication of the locations and degree to which sites are free to vary. For example, some sites may be variable but constrained to vary within pyrimidines (C and T), while others may be free to vary across both purine and pyrimidine nucleotides (A, G, C and T). Multiple alignments based on relatively few sequences can provide a surprisingly good indication of the sequence variation that might exist in nature through combinatoric permutation of the observed sequence variation. Consider the two following aligned sequences, variable at positions 1, 4 and 8:

```
12345678
GAAGCTTA
||||||||
CAACCTTG
```

There are eight ways (2^3) that the observed differences between the two sequences might be permuted:

```
12345678
GAAGCTTA
GAAGCTTG
GAACCTTA
GAACCTTG
CAAGCTTA
CAAGCTTG
CAACCTTA
CAACCTTG
```

The number of combinations implied to be possible expands exponentially as the number of variable sites increases. When this kind of combinatorial expansion is applied to variation that is typical for real data sets, a large number of combinations can result. For example, a pairwise alignment of sequences that vary at 10 sites yields 1024 (2^{10}) combinations. A multiple alignment of eight sequences for which 10 sites are variable for all four nucleotides (A, C, G and T) yields 1 048 576 (4^{10}) combinations. A multiple alignment of eight sequences that shows sequence variation restricted to pyrimidine nucleotides at four sites and among all four different nucleotides at six sites would yield 65 536 ($2^4 \times 4^6$) combinations. The number of combinatorial variants implied by an alignment of a typical gene 1000 bp in length for which only 10% of sites are variable would be far too large to be represented on a chip. However, if the same multiple alignment is broken down into sections of approximately equal length, say between 15 and 30 nucleotides in length, and the implied variation associated with each particular section is computed separately, the number of variants needed to cover the variation implied by the entire sequence alignment is reduced considerably. By arranging oligonucleotides in a series of columns, each of which corresponds to variation in one section of a gene, we are able to circumvent the sequence reassembly problems that are encountered by many alternative methods (11,12). This is the essence of our approach.

Training set

A multiple alignment training set for the protein-coding gene of interest is chosen. Ideally, the multiple alignment should include a phylogenetically balanced sampling of species that is slightly larger in phylogenetic scope than the diversity of the group for which the chip is being designed. This helps to ensure that most of the variants to be encountered in the test data set are likely to be represented in the training data set. The multiple alignment is broken into approximately equal sized sections of between 15 and 30 nucleotides long. All possible sequence variants are computed for each section. This yields the set of probes used to interrogate the unknown (target) sequence. The target sequence is deduced by stringing together the probes that bind the target DNA end to end in the order in which they occur on the multiple alignment.

Oligonucleotide size optimization

For sequencing by hybridization to work reliably, the length of oligonucleotide probes must lie within a narrow range (generally between 15 and 30 nucleotides long) (12). Sequences that are shorter than this optimal range bind weakly, while sequences that are longer are prone to false-positive annealing. Furthermore, as fragment length decreases, there is an increased likelihood that the sequence of the short fragment will occur more than once in the target sequence. This can cause ambiguity when reconstructing sequences from the pattern of bound probes. Highly variable regions such as 'hot spots' will produce more variants than conserved regions. In order to minimize the number of oligonucleotides required to capture the sequence variation implied by a training set, we use a mixture of longer probes to cover the variation for conserved regions and shorter probes to cover the variable regions. We have implemented an algorithm that takes into account the optimal range of probe lengths to cover both conserved and variable regions.

Amino acid filter

When carried out at the nucleotide level, the exponential strategy outlined generally results in an overestimation of the number of sequences that map to a particular protein-coding gene. Some implied combinations, for example, might result in a stop codon in the middle of a coding region, or in the replacement of a structurally critical amino acid at a particular position. This arises because nucleotides are not the primary unit of selection in a DNA sequence. Selection pressure constraining the mutation patterns generally acts at a higher level of abstraction. Our current knowledge does not allow us to specify the exact level (or levels) at which selection acts, but a reasonable first step is to consider the amino acid level. We have implemented a filtering feature that excludes any combination of nucleotides that result in an amino acid that was not present at a particular position in the training set. This results in a marked decrease in the number of oligonucleotide probes required by the chip.

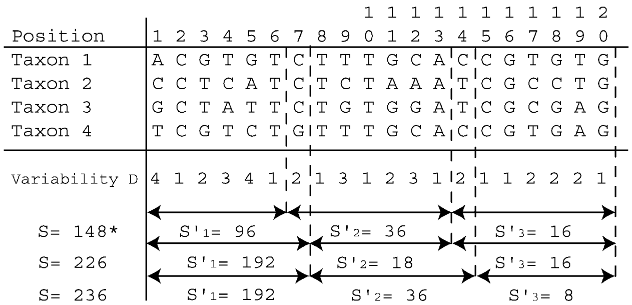
THE ALGORITHM

There are five parts to the algorithm: (i) documenting variability; (ii) computing permutations based on observed variability; (iii) filtering at the amino acid level; (iv) optimizing the length of oligonucleotides; and (v) computing the probes.

(i) A DNA multiple sequence alignment is broken into contiguous blocks. The length of the blocks ranges between an upper and lower bound specified by the user. This range corresponds to the range of probe lengths to be used on the chip and is generally between 15 and 30 nucleotides in length. There are multiple partitioning schemes or 'ways' that a multiple sequence alignment can be divided into non-empty disjoint blocks that completely cover the set. Each partitioning scheme is termed a 'partition', while each block within a partition is termed an 'element' of a partition (Fig. 1). A variability profile is a histogram documenting the number of different nucleotide types present at each site in a multiple alignment (e.g. 1–4). We determine the variability profile for every element j of a partition.

(ii) For each element j, we determine all permutations of nucleotides that could occur based on the observed variation. This is the product of the number of variants observed for each site in the element.

$$S'_j = \prod_{i=1}^{n} V_{ij} \qquad \text{(Equation 1)}$$



Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Taxon 1	A	C	G	T	G	T	C	T	T	T	G	C	A	C	C	G	T	G	T	G
Taxon 2	C	C	T	C	A	T	C	T	C	T	A	A	A	T	C	G	C	C	T	G
Taxon 3	G	C	T	A	T	T	C	T	G	T	G	G	A	T	C	G	C	G	A	G
Taxon 4	T	C	G	T	C	T	G	T	T	T	G	C	A	C	C	G	T	G	A	G
Variability D	4	1	2	3	4	1	2	1	3	1	2	3	1	2	1	1	2	2	2	1

Figure 1. Each path can be scored based on its combinatoric outcome using the two equations **1** and **2**. Consider an alignment composed of 20 nucleotides. In this simplistic example, only two probe sizes are allowed: six and seven oligonucleotide lengths. Thus, we can partition the alignment by three different probe length combinations: 6–7–7, 7–6–7 or 7–7–6. For each site, the nucleotide diversity is counted, the score S' is computed for each partition, and the score S is calculated for each candidate path. The path with the smallest score is chosen (*).

where V_{ji} is the variability for a particular site i for a probe with n sites for a partition element j. Each of the computed permutations yields a potential probe sequence to be used on the chip.

Indels (gaps) are considered as character states at this point of the process. They play the same role as A, C, G and T in the combinatoric expansion described above. However, they are removed at a later point in the procedure [see (v)]. In the event that a probe is computed that has multiple indels such that it is shorter than the lower bound cut off for probe length, it is discarded and not considered in subsequent steps of the algorithm.

(iii) A filter eliminates sequence combinations yielding amino acids that are not present in the training-set and then modifies the score S_j'.

(iv) For each partition, the total number of probes S is determined by summing the score S' of every element:

$$S = \sum_{j=1}^{L} \prod_{i=1}^{n} V_{ij} = \sum_{j=1}^{L} S'_j \qquad \text{(Equation 2)}$$

where L is the number of elements of a partition.

An arrangement that yields the lowest number of probes is selected (Fig. 1).

(v) When a best partition is found, the oligonucleotides are computed. At this point, the indels are eliminated from the final output.

Our implementation uses a graph theoretic representation and a graph search algorithm to resolve a minimization problem.

Graph theoretic representation

For a given DNA sequence alignment, a nucleotide variability vector D is calculated. Let a simple weighted directed acyclic graph $G = (V,E)$ or *dag* (13–15) be a representation of all the possible partitions of D using a range of element length (e.g. 20–30 nucleotides long). Edges represent elements of a partition and vertices represent their starting positions on the vector D. We define the source $s \in V$ as the first position on the vector (in-degree = 0) and the sink $t \in V$ as the last position on

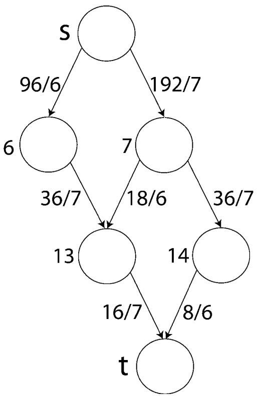


Figure 2. The graph theoretic representation. Each vertex represents a position on the sequence and each directed edge represents an element of a partition [the left value is the weight $w(v_{j-1},v_j) = S_j'$ and the right value is the element length in bp]. The path *s*–6–13–*t* (with probe lengths: 6–7–7) is the path with the smallest score.

the vector (out-degree = 0). A path $p = <s,...,t>$ is a possible partition of D. For instance, for a vector D of length 20 bp, three partitions are possible with elements which are 6 and 7 bp long (Fig. 1). The corresponding *dag* contains four vertices in addition to the source and the sink: vertices representing positions 6, 7, 13 and 14. From the source, it is possible to reach the vertices 6 and 7 by using edges representing elements of respective length 6 and 7 bp; from the vertex 7 we can reach vertices 13 and 14 by using the 6 and 7 bp elements, and so on (Fig. 2). The weight function w: $E \rightarrow Z$ is the total number of probes implied by a path $p = <v_0, v_1,...v_L>$ such as $w(w_{j-1},v_j) = S_j'$ (Equation **1**) and

$$w(p) = \sum_{j=1}^{L} S_j' = S \qquad (\text{Equation } \mathbf{2})$$

For instance, the edge connecting the vertex 6 to the vertex 13 represents a 7 bp element and produces 36 probes; the path from *s* to vertex 13 has $S = 96 + 36 = 132$ probes. The goal is then to find the path $p = <s,...,t>$ that minimizes $w(p)$.

Single-source shortest path algorithm

We use the Dag-Shortest-Path algorithm (15). This algorithm finds the shortest path from vertex *s* to all vertices *v* in linear time, by executing a relaxation of the edges on a topologically sorted *dag*. A topological order of vertices consists of a linear ordering of vertices such that edges are oriented from left to right. A relaxation strategy maintains a value $d[v]$ for each vertex as a shortest path estimate and records the preceding vertex. Relaxation of an edge (v_{j-1},v_j) is carried out if the estimate $d[v_j]$ can be improved {i.e. if $d[v_j] > d[v_{j-1}] + w(v_{j-1},v_{\mathrm{j}})$}. The algorithm proceeds in topological order to all the vertices and all the edges [e.g. (*s*,6), (*s*,7), (6,13), (7,13), (7,14), (13,*t*) and (14,*t*)]. If several partitions are equally weighted, one is chosen based on the order of vertices. The topology sorting is trivial since the *dag* is built 'sequentially'; the order of the vertices corresponds to the position they represent on the vector D (e.g. *s*,6,7,13,14,*t*). Thus, the algorithm requires only one pass over the vertices to complete the search.

PROOF OF CONCEPT

Number of probes necessary to capture variation observed in nature

We tested our approach using a number of different data sets downloaded from GenBank. We selected cytochrome-b gene sequences (average length 1150 bp) from six different vertebrate clades: carnivores, artiodactyls, rodents, marine carnivores, sharks and whales. Cytochrome-b was chosen because it is one of the most commonly used markers for comparative phylogenetics and is well represented in GenBank. We aligned each data set using ClustalW (16). We then computed the number of oligonucleotide variants required to capture the variation implied by the training sets using the algorithm described. When the probe length is set to a constant, say 20 bp, the predicted number of probes is generally much larger than current chip technology will allow (Table 1, row 1). When probe length is allowed to vary between 20 and 30 bp, the number of oligonucleotide variants ranges from 23 914 for the whale data set to 12 146 788 for the primates (Table 1, row 2). While this represents a marked improvement, half of the data sets tested yielded more probes than would fit on a chip. When we used a range of shorter probes (13–15 bp), the number of oligonucleotide variants does not exceed 197 708 (Table 1, row 3). Unfortunately, such short probe lengths would probably present technical difficulties related to hybridization specificity. Allowing variation in

Table 1. Summary of the different groups tested by simulation for the accuracy of the chip

	No. of variants							
Probe length	Carnivores	Sharks	Whales/dolphins	Artiodactyls	Marine carnivores	Rodents/lagomorphs	Primates	Fish
20 bp	2 329 184	234 840	55 770	4 683 388	113 160	1 866 616	16 185 212	9 708 908
20–30 bp	1 219 452	151 228	23 914	1 933 940	51 980	966 396	13 146 788	6 294 968
13–15 bp	45 366	12 338	4421	62 404	8200	41 358	197 708	97 947
15–25 bp	103 366	23 346	6858	146 786	13 808	89 961	592 272	299 562
18–22 bp	486 596	72 992	15 158	663 968	30 444	313 592	3 771 860	1 541 024
18–25 bp	484 260	72 992	15 134	663 392	30 204	312 720	3 769 400	1 535 536
No. of taxa	24	13	14	25	16	12	22	18
Accuracy (%)	98.28	98.77	98.99	99.26	98.96	97.92	98.76	97.60

Note that the 20–30 bp range may be technically unrealistic because of melting temperature problems, as well as the 13–15 bp range for hybridization specificity; this example is only used to estimate the algorithm efficiency.
The percentage accuracy represents the percentage of nucleotides recovered by the chip.

probe length significantly reduces the number of variants required relative to a chip of fixed probe length (Table 1, rows 4–6) because probe length can be tailored to accommodate local fluctuations in variability (shorter probes are used for variable regions while longer probes are used for more conserved regions). The choice of the range of probe lengths is a trade-off between the hybridization limitations and the chip size limitations. Interestingly, our results show little correlation with the number of species (i.e. sequences) used in the training set, but, as might be expected, show the effect of factors such as degree of evolutionary divergence among the training set group (Table 1, rows 7 and 8).

Testing the combinatorial efficacy of the chip

To test the efficacy of our 'virtual chips' (i.e. suite of estimated probes associated with each training set), we adopted a jacknifing procedure in which each species was removed in turn from the training set and considered as the target sequence to be determined. We scored the number of unknown sequences correctly identified by this procedure and define as our measure of accuracy the percentage of nucleotides correctly identified (Table 1, column 5). On average, we attain an accuracy of ~98% for the test data sets used. While this gives a coarse sense of the combinatorial coverage attained by the chip, it should not be taken to mean that the procedure would identify 98% of the nucleotides of most sequences. Rather it means the procedure will identify most sequences perfectly and a few very poorly. This is because a failure to hybridize associated with a single nucleotide mismatch results in the loss of data for an entire probe rather than a single mismatching nucleotide. When a probe fails to bind, the whole set of nucleotides associated with the probe is undetermined. Our tests also indicate that synonymous changes are responsible for most of the mismatches (data not shown).

DISCUSSION

Initial results indicate that the approach is sensitive to the sampling in the training set. There are two aspects to sampling: (i) the number of sequences and (ii) the 'quality' of the sampling. A chip is best designed from a training set of sequences that contain a diverse and representative sampling of taxa. Only a few carefully selected taxa are required to cover the non-synonymous class of changes. In contrast, the accuracy of synonymous changes will continue to improve as more taxa are added. It is important to note that while combinatorial expansion can result in a large number of implied variants, it can occasionally miss some of the variants that might occur in nature. This happens when some of the variation that occurs in nature is not captured by the variation implied by the training set. As in any sampling problem, the more representative the training set, the more effective the combinatorial coverage.

Our approach highlights several problems inherent to the structure of the data itself. Genes with hotspots require higher variance of probe lengths than genes whose sequence variation is more evenly distributed. Repetitive DNA increases the likelihood that probes in different areas of the alignment would have the same sequence causing ambiguity in the sequence reconstruction phase. Sequences with high G–C contents can generate melting temperature heterogeneity issues (this could be circumvented by including a filter in the algorithm that eliminated probes with undesirable melting temperatures in the same way that probes containing certain amino acids are currently eliminated). Finally, it should be noted that the approach requires that the multiple alignment itself be reliable. If the alignment is incorrect, the estimated probes may fail to bind the target sequence.

The gene-specific chip design we propose will obviously not be effective in all situations. The approach assumes that the training data set captures most of the variability for the studied gene. There will undoubtedly be cases where a chip is applied to a target whose sequence is not covered by the variation implied by the training set. However, the design is such that we will know when this occurs. The target will simply fail to bind a probe associated with the region in question. Such negative data will show up as a lack of fluorescence. In such circumstances, the target template can be identified and sequenced by conventional means. The method should be well suited to training sets with a low level of diversity. We envisage particular utility for population genetics, gene-specific disease screening, medical diagnostics, forensic and epidemiological applications. The approach would be especially efficient for screening mutations associated with diseases such as breast and ovarian cancer within the BRCA1 exon 11. Current approaches use arrays that are 'hardwired' to detect specific mutations. For example, Hacia *et al.* (17) used such a chip based on 96 000 probes, and a diagnostic chip using 400 000 probes is under development (18). Our approach would allow for a more flexible screening of combinations of mutations that while not observed directly in the training set, are implied by the training set to be possible in nature.

We see opportunities for further reducing the number of probes required by incorporating filters that use higher order features such as secondary and tertiary structural considerations. As more is learned about the mapping between sequence variation and protein structure (and function), and about constraints at the level of the genetic code, the number of oligonucleotide probes can be reduced.

The perfectly optimized chip will completely describe the evolutionary 'opportunity space' associated with the protein to be sequenced (19), and will reflect the protein's function. Understanding where the boundaries of a particular protein's opportunity space lie will improve our understanding of the interaction between the molecular evolutionary process and the mapping between underlying genotype and protein phenotypes. We see the presented chip design as providing a framework for high throughput screening of many thousands of templates and for identifying the minority of novel templates that require sequencing by other means.

ACKNOWLEDGEMENTS

We thank Vasant Honavar and Oliver Eulenstein in the Department of Computer Science at Iowa State University; Vasant Honavar for his input and advice in the early phase of the project and Oliver Eulenstein for his careful review and for pointing out that our original solution to the problem was a *dag* and a more efficient algorithm could be used to solve it.

We thank Anna Keyte for editing various versions of the manuscript.

REFERENCES

1. Hunkapiller,T., Kaiser,R.J., Koop,B.F. and Hood,L. (1991) Large-scale and automated DNA sequence determination. *Science*, **254**, 59–67.
2. Bains,W. and Smith,G.C. (1998) A novel method for nucleic acid sequence determination. *J. Theor. Biol.*, **135**, 303–307.
3. Lysov,Y.P., Florentiev,V.L., Khorlon,A.A. Khrapko,K.R., Shik,V.V. and Mirzakebov,A.D. (1988) Sequencing by hybridization via oligonucleotides. *Dokl. Acad. Nauk SSSR*, **303**, 1508–1511.
4. Drmanac,R., Labat,I., Brukner,I. and Crkvenjakov,R. (1989) Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics*, **4**, 114–128.
5. Southern,E.M. (1996) DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. *Trends Genet.*, **12**, 110–115.
6. Wallace,R.B., Shaffer,J., Murphy,R.E., Bonner,J., Hiros,T. and Itakura,K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to X174 DNA: the effect of a single base pair mismatch. *Nucleic Acids Res.*, **6**, 3543–3557.
7. Thein,S.L. and Wallace,R.B. (1986) The use of synthetic oligonucleotides as specific hybridization probes in the diagnosis of genetic disorders. In Davies,E.K. (ed.), *Human Genetic Disease: A Practical Approach*. IRL Press, Oxford, UK, pp. 33–50.
8. Wallraff,G., Labadie,J., Brock,P., Dipietro,R., Nguyen,T., Huynh,T., Hinsberg,W. and McGall,G. (1997) DNA sequencing on a chip. *Chemtech*, **27**, 22–32.
9. Preparata,F.P., Fireze,A.M. and Upfal,E. (1999) On the power of universal bases in sequencing by hybridization. In *Recomb'99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*. Lyon, France, pp. 295–301.
10. Pevzner,P.A., Lysov,Y.P., Khrapko,K.R., Belvavsky,A.V., Florentiev,V.L. and Mirzabekov,A.D. (1991) Improved chips for sequencing by hybridization. *J. Biomol. Struct. Dynam.*, **9**, 399–410.
11. Winzeler,E.A., Richards,D.R., Conway,A.R., Goldstein,A.L., Kalman,S., McCullough,M.J., McCusker,J.H., Stevens,D.A., Wodicka,L., Lockhart,D.J. and Davis,R.W. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.
12. Southern,E.M., Maskos,U. and Elder,J.K. (1992) Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics*, **13**, 1008–1017.
13. Sakarovitch,M. (1984) *Optimisation Combinatoire: Méthodes Mathématiques et Algorithmiques*. Collection Enseignement des Sciences, Hermann, Paris.
14. Cook,W.J., Cunningham,W.H., Pulleyblank,W.R. and Schrijver,A. (1998) *Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York.
15. Cormen,T.H., Leiserson,C.E. and Rivest,R.L. (1997) *Introduction to Algorithms*. MIT Press. McGraw-Hill Book Co.
16. Thomson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensivity of progressive multiple alignment through sequence weighting, position, gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. Hacia,J., Brody,L., Chee,M., Fodor,S. and Collins,F. (1996) Detection of heterozygous mutation in BRCA1 using high density oligonucleotide array and two-colour fluorescence analysis. *Nature Genet.*, **14**, 441–447.
18. Ramsay,G. (1998) DNA chips: state-of-the art. *Nat. Biotechnol.*, **16**, 40–44.
19. Naylor,G.J.P. and Gerstein,M. (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. *J. Mol. Evol.*, **51**, 223–233.