Sequencing by Hybridization with Positive Faults

Jacek Błażewicz [1] blazewic@sol.put.poznan.pl
Piotr Formanowicz [1] piotr@cs.put.poznan.pl
Marta Kasprzak [1] marta@cs.put.poznan.pl
Wojciech T. Markiewicz [2] markwt@ibch.poznan.pl
Jan Węglarz [1 2] weglarz@sol.put.poznan.pl

[1] Institute of Computing Science, Poznan University of Technology, Piotrowo 3A, 60-965 Poznan, Poland

[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

Abstract

The paper is concerned with a computational phase of the sequencing DNA chains by hybridization. It is assumed that positive faults can occur in the hybridization experiment. An approach based on a reduction of the problem to a variant of a Selective Traveling Salesman Problem and an algorithm for solving the latter, have been proposed. The algorithm behaves extremely well, even for a fault rate exceeding 50%.

1 Introduction

One of the most challenging issues in the area of molecular biology is to read (recognize) a structure of a human genome (that is a DNA sequence). DNA sequencing aims at discovering the exact sequence of nucleotides in rather short DNA chains (100-400 nucleotides long). (Longer sequences are assembled with the use of different approaches, cf. [23].) One of the variants of this method is sequencing by hybridization (SBH), proposed among the others by Drmanac et al. [8], Khrapko et al. [18] Southern et al. [27] and Markiewicz et al. [21].

In general, the hybridization experiment is based on the property of single stranded nucleic acids to form a complex (antiparallel) with a complementary strand of a nucleic acid [5, 25, 28]. All short fragments of nucleic acids (oligonucleotides) of length l are used in the hybridization experiment (so called library of 4^l different oligonucleotides) and thus, the formation of the complex indicates the occurrence of a sequence complementary to the oligonucleotide in the nucleic acid. It is detected by a nuclear or spectroscopic detector. As a result of the experiment one gets a set (called spectrum) of all l-long oligonucleotides which are known to hybridize with the investigated DNA sequence N of length n (i.e., they are substrings of string N). In a case of ideal data we have thus $\mid spectrum \mid = n-l+1$. In some cases, however, one can obtain less than $n-l+1$ fragments, either due to experimental problems or as a result of the structure of sequence N (for example, it can contain repeated subsequences). In this case negative faults occur. On the other hand, there can be also positive faults, i.e. not each fragment (oligonucleotide) present in spectrum is a part of the original sequence.

Due to the differences in the thermodynamic stability of perfectly matched duplexes of the same length l, resulting in the first instance from differences in a A/T, G/C content, it is not possible to perform an ideal hybridization experiment, in general. The addition of chaotropic salts into the hybridization mixtures allows to reduce this source of faults only to some extent [22]. However, it is possible to perform hybridization experiment under such conditions (i.e. less stringent conditions) that all l long oligonucleotides of probing library (chip) whose complementary oligonucleotides are present in the DNA sequence N, even the one forming the least stable duplex, will be detected. On the other hand, hybridization experiments should be performed in a way which will result in a minimum number of experimental faults. One might assume that a hybridization experiment should supply a data set which contains less faulty than correct l-mers and thus, in sequence N reconstruction procedure,

a maximum number of hybridizing l-mers should be used. (If a solid supported full combinatorial library with all possible l long oligonucleotides is used in the hybridization experiment, then negative experimental faults are excluded). This means that the negative faults of the experimental nature can thus be avoided. For DNA sequences in which repeated subsequences of length larger than l are not present, this allows to eliminate all negative faults. In the same time, performing a hybridization experiment under stringent conditions will certainly result in positive faults.

A computational phase of the SBH method (i.e. a reconstruction of a DNA chain on the basis of the obtained spectrum) can be done in polynomial time [24] in case of an ideal spectrum. More difficult (in fact NP-hard) [12, 13] are the cases where some faults occur. Several heuristics have been proposed for the case where only negative faults appeared [1, 2, 4, 15, 18, 24]. Only a few papers have dealt with a limited case of both positive and negative faults [9, 19], assuming that only the first or the last bit (nucleotide) in the obtained data can be wrong. In the recent paper [3] a new method has been proposed that takes into account both types of faults without any assumption on their type and source. Its computational behavior is satisfactory for the fault rate achieving 10%.

As was discussed above, it is possible to tune the hybridization experiment in such a way that the negative experimental faults can be avoided. In the present paper, we use this assumption to modify the general method presented in [3]. It appears that in the case of only positive faults, the method performs extremely well, even for the fault rate exceeding 50%.

In the next Section the method and its modification will be described. Section 3 will contain results of the computational experiments.

2 The Method

As we discussed, the problem to be considered may be formulated as follows. Given *spectrum* (a set of l-mers that hybridized with an analyzed DNA sequence N) and length n of sequence N, reconstruct the latter. The *spectrum* may contain positive faults. Our approach is based on an idea that when reconstructing a DNA sequence, a maximum cardinality subset of l-mers from *spectrum* should be used. Moreover, a number of nucleotides in the obtained sequence should not be greater than in the source sequence. Joining two l-mers into a sequence is connected with a cost. Two l-mers of length l may overlap on $l-1, l-2, \ldots,$ or 0 nucleotides, respectively. The cost of joining two l-mers is equal to l minus a number of nucleotides that overlap in these l-mers. For example, two l-mers CCAGA and GATTC may overlap on two nucleotides and create a longer sequence CCAGATTC. Thus, a cost of joining them is equal to 3. A sequence of length n obtained by joining l-mers which overlap with their neighbors on $l-1$ nucleotides, contains $n-l+1$ l-mers. Usually l-mers being positive faults cannot be overlapped on both their ends with proper l-mers from the *spectrum*.

We propose to formulate the problem of constructing a DNA sequence from a *spectrum* containing positive faults as a variant of a *Selective Traveling Salesman Problem* (STSP). In a classical *Traveling Salesman Problem* (TSP) there is given a complete directed or undirected graph where to each arc or edge a cost is assigned. The goal is to visit every vertex in the graph exactly once and return to the starting point in such a way that a sum of costs of traversed arcs or edges included in the directed cycle is at its minimum. This problem is known to be strongly NP-hard, thus unlikely to admit a polynomial-time algorithm [14].

The STSP is a modification of the classical version. In this problem in addition to arc costs there is defined a profit for each vertex. The goal is to visit as many vertices as possible. In other words, a sum of profits of visited arcs is maximum and the total cost of traversed edges does not exceed a given value. In this version of the problem the first vertex must be also the last one.

If in the STSP the path is not required to be a cycle and each vertex profit is equal to 1, then the new problem is equivalent to the DNA sequencing problem in the presence of positive and negative faults. The method solving this problem has been presented in [3]. In this case each l-mer from the *spectrum* corresponds to a vertex in the graph. Each arc incident with two vertices in the graph is

labeled by an overlapping cost of the two l-mers corresponding to these vertices. One should notice that this cost depends on the order in which these two l-mers are joined. Hence the graph must be directed. Moreover, there are no loops in this graph.

If in the above graph arcs corresponding to connections of l-mers overlapping on less than $l-1$ nucleotides (i.e. which costs are greater than 1) have been removed from the graph, then this version of STSP is equivalent to the DNA sequencing problem with only positive faults. Notice that this graph is very sparse and a corresponding search tree consists of a small number of nodes. Computing an effective upper bound in such a tree would be very time consuming. An eventual profit (a number of cut off nodes) obtained by using this bound would not justify computing it.

Note, that in a case when a path is a cycle this method can be used to sequence circular DNA strands. We assume that l-mers are always maximally overlapped. For example, if there are two l-mers: ACTTC and CTTCG then they overlap on four nucleotides and create longer sequence ACTTCG with cost equal to 1.

Constructing a desired path in the formulated above variant of *Selective Traveling Salesman Problem* is equivalent to finding a subset of cardinality $n-l+1$ of l-mers from *spectrum* that hybridized with unknown sequence N. Note that in this particular case the value of an optimal solution is known (and equal to $n-l+1$). Thus, this modified version of STSP may be formulated as the following system of linear equations and inequalities:

$$\sum_{i=1}^{z}\sum_{j=1}^{z} x_{ij} = n-l \tag{1}$$

$$\sum_{i=1}^{z} x_{1i} = 1 \tag{2}$$

$$\sum_{i=1}^{z} x_{i1} = 0 \tag{3}$$

$$\sum_{i=1}^{z} x_{ki} \leq 1,\ k = 2,\ldots,z \tag{4}$$

$$\sum_{i=1}^{z} x_{ik} \leq 1,\ k = 2,\ldots,z \tag{5}$$

$$\sum_{i=1}^{z} x_{ki} - \sum_{j=1}^{z} x_{jk} \leq 0,\ k = 2,\ldots,z \tag{6}$$

$$\sum_{v_k \in S} (\sum_{i=1}^{z} x_{ik} + \sum_{j=1}^{z} x_{kj}) \leq 2|S| \sum_{v_i \in V\setminus S, v_j \in S} x_{ij},\quad S \subset V \setminus \{v_1\} \tag{7}$$

where:

z - a cardinality of *spectrum*; its value is equal to or greater than $n-l+1$,

V - a set of vertices (i.e. l-mers) of cardinality z,

x_{ij} - a boolean variable; it is equal to 1 if an arc joining vertices i and j is included in the solution; otherwise it is equal to 0.

The first vertex (i.e. l-mer) of the path is assumed to be known. (Later we will show how to omit this condition.) In the above formulation it is denoted by variables with index 1.

In the *spectrum* there are no negative faults hence a number of vertices selected in the traveling salesman path (i.e. l-mers constituting sequence N) is equal to $n - l + 1$. This condition is expressed by equation (1). It is obvious that each l-mer in a solution, except for a starting one and an ending one, must have exactly one immediate predecessor and exactly one immediate successor. Moreover, the starting l-mer must have one immediate successor but no predecessors. Similarly, the ending l-mer must have one immediate predecessor but no successors. Equation (2) guarantees that the starting vertex has only one immediate successor in the solution. Similarly, equation (3) guarantees that in the solution there is no predecessor of the starting vertex. Inequalities (4) and (5) ensure that each vertex included in the solution, except for the starting one, has at most one immediate predecessor and one immediate successor in the solution path. According to inequalities (6) there is only one starting point in the path. Inequality (7) ensures that there is no separate cycle in the solution. S is any subset of set $V \setminus \{v_1\}$. The left side of this inequality is a sum of degrees of all vertices in S. From (4) and (5) it follows that this sum may take values at most equal to $2|S|$. The right hand side of this formula is a number of all arcs which have a starting vertex in $V \setminus S$ and an ending one in S, multiplied by $2|S|$. Hence, inequality (7) does not hold only in the case where there is no such arc. If (2) - (6) hold, this situation would take place only if there has been a separate cycle in the graph.

To solve the above problem an algorithm based on the branch and bound approach is proposed.

The required length of sequence (n), the set of oligonucleotides (*spectrum*) and one element from *spectrum* assumed to be an initial one, are known at the beginning of a searching process. One should obtain final sequences, each of them of length equal to n, consisting of $n - l + 1$ oligonucleotides from *spectrum*. A search tree, which is built during this process, has elements of *spectrum* (l-mers - words of length l) as nodes. A path consisting of visited nodes in the tree corresponds to an arrangement of l-mers included into the sequence. The path begins with an initial oligonucleotide (l-mer) and is built until it becomes an acceptable solution, or until further branching is not feasible. The algorithm described in this section does not need any knowledge of an initial l-mer. If such an l-mer is not indicated, the algorithm successively treats each oligonucleotide as an initial one. In this case each of the l-mers becomes a root of a subtree built by the algorithm.

At each node the search tree is branched. A set of successors of any node is the same as in the graph with an exception of vertices already visited. If during the search process a level $n - l + 1$ in the search tree is achieved, a current path is remembered as a solution.

3 Computational Results

The described algorithm has been tested on DNA sequences obtained from GenBank, National Institute of Health, USA. Sequences coding human proteins have been chosen. Computational experiments have been carried out on SGI Power Challenge in Poznan Supercomputing and Networking Center.

For these tests 10 DNA strings have been selected, for which there were no repetitions of 10-mers among the first 409 nucleotides. In order to obtain *spectra* consisting of 100, 200, 300 and 400 elements respectively, 109, 209, 309 and 409 initial nucleotides from these strings have been taken. Next, a hybridization experiment has been simulated by cutting off oligonucleotides of length 10 from the long sequence. The lengths of long strings and short oligonucleotides have been chosen on the basis of real biochemical experiments [6]. The tests described in [3] have shown that in majority of real DNA chains of length between 100 and 400 nucleotides there are no natural repetitions of oligonucleotides of length 10. In such a case it is possible to tune the hybridization experiment in such a way that the negative experimental faults can be avoided. Because of this the described algorithm is useful in practice. In cases where such repetitions appear the most general algorithm described in [3] should be used. One should notice that for strands with positive faults only the specialized algorithm

outperforms the general one. Moreover oligonucleotides of length 10 give high probability that there is only one solution. This fact has been confirmed by tests - in all cases one solution, identical with a source sequence, has been obtained. But our algorithm has been also adapted to any values of n and l ($l \leq n$) and always generates a full list of feasible solutions. In tests a starting l-mer has been known. A notation "100+20" means that into an ideal *spectrum* (without any faults) of cardinality 100, 20 randomly generated positive faults have been added. An addition of next 20 l-mers created a set "100+40" etc. l-mers which have been included had to be different from those already existing in the *spectrum*.

Below results of the computational experiment for the described algorithm are presented. Computation times of the algorithm accepting all types of faults (described in [3]), are also shown. They indicate a speed-up obtained by the algorithm constructed for the case where only positive faults appear.

Table 1 contains average computation times (10 runs for each entry) depending on a length of a source sequence and a number of faults. The computation times have been restricted to 500 seconds and average times shown in Table 1 take into account only those instances which have been solved in the time limit. Table 2 depicts a number of such instances.

A number of positive faults in *spectrum*	The algorithm for positive faults only				The algorithm for all types of faults			
	A number of l-mers in a source sequence							
	100	200	300	400	100	200	300	400
0	0.02	0.05	0.11	0.20	0.64	30.22	21.05	82.61
20	0.02	0.06	0.13	0.21	1.16	42.13	28.07	96.47
40	0.03	0.07	0.14	0.24	1.97	55.64	35.73	64.61
60	0.03	0.08	0.16	0.26	3.25	9.98	43.40	67.94
80	0.04	0.10	0.18	0.28	4.71	12.36	52.66	86.22

Table 1. Average computation times (in seconds) for, respectively, the algorithm accepting positive faults only, and the algorithm accepting all types of faults (instances with positive faults only).

A number of positive faults in *spectrum*	The algorithm for positive faults only				The algorithm for all types of faults			
	A number of l-mers in a source sequence							
	100	200	300	400	100	200	300	400
0	10	10	10	10	10	10	9	9
20	10	10	10	10	10	10	9	9
40	10	10	10	10	10	10	9	8
60	10	10	10	10	10	9	9	8
80	10	10	10	10	10	9	9	8

Table 2. A number of instances (from among 10 used in the experiment) solved by the algorithms within 500 seconds.

Tests of the presented algorithm have given surprisingly good results, even for 80% introduced faults (as compared with an ideal *spectrum*), hence we have decided to replace completely randomly generated errors by more realistic ones. In order to include one false element into a *spectrum*, one of the l-mers already existing in it has been duplicated and then its first, last or both nucleotides have been changed. Because for this kind of faults results were almost identical to those obtained earlier (the difference was not greater than 0.01 second), we have decided to prepare more complicated input data. The worst type of faults for the described algorithm is the one where the last nucleotide of an l-mer is false. A series of tests with only this type of faults have been carried out. Again results were very similar to those obtained for the randomly generated faults (Table 1).

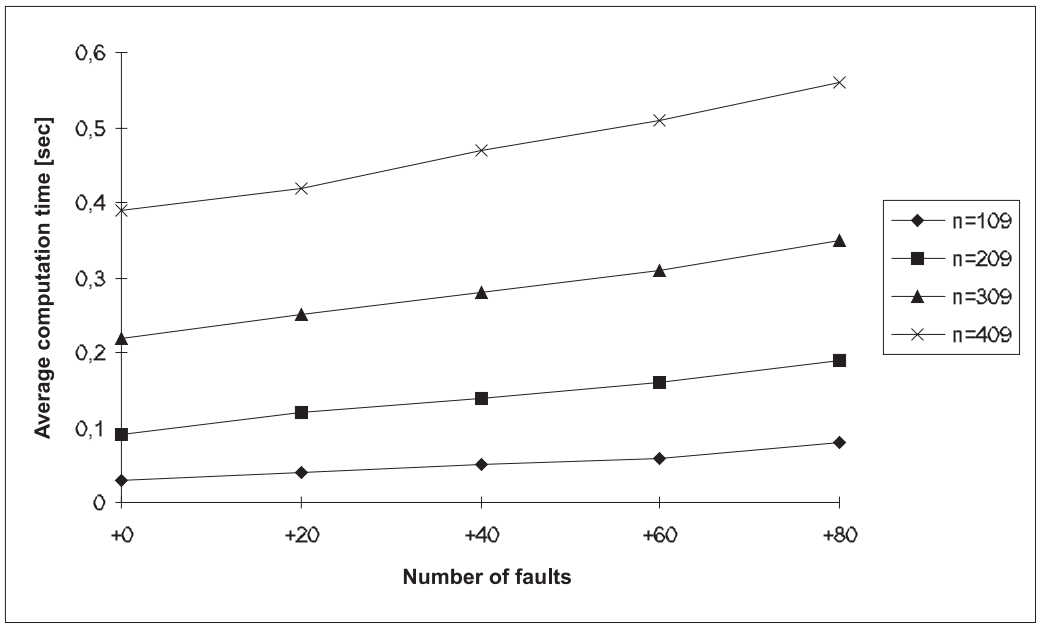

Figure 1: Figure 1. An average computation time (in seconds) versus the number of positive faults in *spectrum* (for various lengths n of an original sequence) for the algorithm accepting positive faults only

4 Conclusions

In the paper, the algorithm for solving a computational phase of the DNA sequencing by hybridization method, has been proposed. The algorithm can handle the case positive faults, appearing in the data obtained from the hybridization experiment. No a priori knowledge about the source of these faults is required.

The presented algorithm has been extensively tested on real, thus computationally harder, DNA sequences. Parameters n (the length of a reconstructed sequence) and l (the length of hybridizing oligonucleotides) chosen for tests purposes, had the values used in real experiments. The algorithm has behaved exceptionally well if only positive errors have appeared (and this knowledge could be used). The algorithm can reconstruct DNA sequences from *spectra* containing of up to 50% (positive) faults, in a very short time.

Future examinations should allow for a further evaluation of the scope of applications of the sequencing approach with positive faults described in this paper.

References

[1] Bains, W. Hybridization methods for DNA sequencing, *Genomics* 11:294–301, 1991.

[2] Bains, W., Smith, G.C. A novel method for nucleic acid sequence determination, *J. Theor. Biol.* 135:303–307, 1988.

[3] Blazewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., Weglarz, J. DNA Sequencing with Positive and Negative Errors. Report, Poznan Supercomputing and Networking Center, Poznan 1997 (submitted for publication).

[4] Blazewicz, J., Kaczmarek, J., Kasprzak, M., Markiewicz, W.T., Weglarz, J. Sequential and parallel algorithms for DNA sequencing, *CABIOS*, vol. 13 no. 2:151–158, 1997.

[5] Blazewicz, J., Kasprzak, M., Sterna, M., Weglarz, J. Selected combinatorial optimization problems arising in molecular biology, *Ricerca Operativa*, 1997, to appear.

[6] Caviani Pease, A., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., Fodor, S.P.A. Light-generated oligonucleotide arrays for rapid DNA sequence analysis, *Proc. Natl. Acad. Sci. USA*, 91:5022–5026.3, 1994.

[7] Chee, M., Cronin, M.T., Fodor, S.P.A., Gingeras, T.R., Huang, X.C., Hubbell, E.A., Lipshutz, R.J., Lobban, P.E., Miyada, C.G., Morris, M.S., Shah, N., Sheldon, E.L., and Fodor, S.P. New arrays of oligonucleotide probes - used for comparing known sequences with variants for detection of mutation(s) and sequencing. Affymax Technologies NV. USA:WO 9511995, 1995.

[8] Drmanac, R., Labat, I., Brukner, I., Crkvenjakov, R. Sequencing of megabase plus DNA by hybridization: theory and the method, *Genomics* 4:114–128, 1989.

[9] Drmanac, R., Labat, I., Crkvenjakov, R. An Algorithm for the DNA Sequence Generation from k-Tuple Word Contents of the Minimal Number of Random Fragments, *Journal of Biomolecular Structure & Dynamics*, vol. 8, no. 5:1085–1102, 1991.

[10] Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. Light-Directed, Spatially Addressable Parallel Chemical Synthesis. *Science* 251:767–773, 1991.

[11] Fodor, S.P.A., Goldberg, M.J., Goss, V., Mcgall, G., Pease, R.F., Rava, R.P., Stryer, L., and Winkler, J.L. Forming polymers having diverse monomer sequences on substrate - where substrate comprises linker mol. and protective gp., applying barrier layer and exposing regions of linker mol. layer to vapour comprising deprotecting agent. Affymax Technologies NV.EP 728520, 1996.

[12] Gallant, J., Maier, D., Storer, J.A. On Finding Minimal Length Superstrings, *Journal of Computer and System Sciences* 20:50–58, 1980.

[13] Gallant J.K. The Complexity of the Overlap Method for Sequencing Biopolymers, *J. Theor. Biol.* 101:1–17, 1983.

[14] Garey, M.R., Johnson, D.S. Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Co. 1979.

[15] Guenoche, A. Can we recover a sequence, just knowing all its subsequences of given length?, *CABIOS* 8:569–574, 1992.

[16] Hacia, J.G., Brody, L.C., Chee, M.S., Fodor, S.P.A., and Collins, F.S. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nature Genet.* 14(4):441–447, 1996.

[17] Khorlin, K.K., Khrapko, K.R., Lysov, Y.P., Yershov, G.K., Vasilenko, S.K., Florentiev, V.L., and Mirzabekov, A.D. An oligonucleotide matrix hybridization approach to DNA sequencing. *Nucleic Acids Res.*, Symp.Ser. 24:191–192, 1991.

[18] Khrapko, K.R., Lysov, Y.P., Khorlin, A.A., Shik, V.V., Florent'ev, V.L., Mirzabekov, A.D. An oligonucleotide approach to DNA sequencing, *PEBS Letters* 256: 118–122, 1989.

[19] Lipshutz, R.J. Likelihood DNA Sequencing By Hybridization, *Journal of Biomolecular Structure & Dynamics*, vol. 11, no. 3:637–653, 1993.

[20] Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., Shen, N., Yang, R., and Fodor, S.P.A. Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques* 19(3):442–447, 1995.

[21] Markiewicz, W.T., Andrych-Rozek, K., Markiewicz, M., Zebrowska, A., Astriab, A. Synthesis of oligonucleotides permanently linked with solid supports for use as synthetic oligonucleotide combinatorial libraries. Innovations in Solid Phase Synthesis, in: *Biological and Biomedical Applications* (R. Epton, ed.), Mayflower Worldwide: 339–346, 1994.

[22] Maskos, U. and Southern, E.M. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support. *Nucleic Acids Research* 21:4663–4669, 1993.

[23] Myers, E.W. Toward Simplifying and Accurately Formulating Fragment Assembly, *J. Comput. Biol.*, vol. 2, no. 2:275–290, 1995.

[24] Pevzner, P.A. l-Tuple DNA Sequencing: Computer Analysis, *Journal of Biomolecular Structure & Dynamics*, vol. 7, no. 1:63–73, 1989.

[25] Pevzner, P.A., Waterman M.S., Open Combinatorial Problems in Computational Molecular Biology, Proceedings of Israel Symposium on the Theory of Computing and Systems, Tel Aviv, Israel, January 4–6, 158–173, 1995.

[26] Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H. Primer-Directed Enzymatic Amplification of DNA with Thermostable DNA Polymerase. *Science* 239:487–491, 1988.

[27] Southern, E.M., Maskos, U., and Elder, J.K. Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation Using Experimental Models. *Genomics* 13:1008–1017, 1992.

[28] Waterman, M.S. Introduction to Computational Biology. Maps, sequences and genomes, Chapman & Hall, 1995.