Isothermic Oligonucleotide Libraries

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

[1] Institute of Computing Science, Poznań University of Technology, Piotrowo 3a, 60-965 Poznań, Poland
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland

Keywords: sequencing by hybridization, combinatorial oligonucleotide libraries

1 Introduction

Reading a sequence of an unknown DNA remains one of the most challenging issues in molecular biology. One of the sequencing methods is sequencing by hybridization (SBH), which has been proposed, for example, by Southern et al. [5]. The use of standard oligonucleotide libraries leads to an excessive number of experimental errors, combinatorial methods must deal with (see e.g. [3, 4]).

On the other hand it is well known that DNA duplexes of C/G rich l-mers are more stable than A/T rich ones. This serious obstacle can result in numerous errors of a positive type. In the same time a modification of hybridization conditions directed towards diminishing the number of imperfect duplexes of C/G rich l-mers might result in increasing a number of missing perfect duplexes of A/T rich l-mers. The duplex formation depends on a base composition but also on its length. In the earliest studies using allele specific oligonucleotides in DNA mutation analysis a simple equation was used to calculate melting temperatures of oligonucleotide duplexes assuming 4 degrees for C/G pairs and 2 degrees for A/T pairs [6]. Thus, taking formally, it is possible to compensate lower stability of A/T rich duplexes by increasing their lengths. It is known that the above description is not very accurate although it reflects a general relative stability of different duplexes quite well. However, in our opinion this simple way of calculation of duplex formation taking into account only numbers of A/T and C/G pairs with appropriate increments offers a new solution to a sequencing by hybridization (cf. [1]).

Therefore, we propose to obtain a set of oligonucleotides that differ in base composition and length and are characterized by predefined relations between base composition and a length of oligonucleotides, as defined in Section 2. Sets of such oligonucleotides will be called *isorelational oligonucleotide libraries*. In a specific case if in a library an increment of C(G) is twice of A(T) and the sum of increments for each oligonucleotide is constant then such library is called *isothermic oligonucleotide library*. Their use results in a higher stability of duplexes obtained by SBH, thus, reducing a number of experimental errors.

2 Isothermic Libraries

In the previous section we gave a rough description of isorelational oligonucleotide library. To be more precise define them more formally. *An oligonucleotide library L consisting of all oligonucleotides satisfying relation $w_A x_A + w_C x_C + w_G x_G + w_T x_T = C_L$, where w_A, w_C, w_G, w_T are increments of nucleotides A, C, G and T, respectively, and x_A, x_C, x_G, x_T denote numbers of these nucleotides in the oligonucleotide, and C_L is a constant parameter for the library, is an isorelational oligonucleotide library.*

We see that by taking in addition a very simple formula $D = x_A + x_C + x_G + x_T$, where D is a length of an oligonucleotide, we have also a relation between base composition and a length of the oligonucleotide. Note, that if $w_A = w_C = w_G = w_T$ for a given library, one gets a standard oligonucleotide library.

As a special case of these libraries we have isothermic oligonucleotide libraries, formally defined below. *An isothermic oligonucleotide library L of temperature T_L is a library of all oligonucleotides satisfying relations $w_A x_A + w_C x_C + w_G x_G + w_T x_T = T_L$, $w_A = w_T$, $w_C = w_G$ and $2w_A = w_C$.*

Without loss of generality we assume here that $w_A = w_T = 2$ and $w_C = w_G = 4$. This corresponds to increments particular nucleotides bring into stability of oligonucleotide duplexes. In what follows a sum of increments of nucleotides forming an oligonucleotide will be called an oligonucleotide temperature.

Isothermic oligonucleotide library follows the experimentally established relationship between base composition and duplex stability described in Introduction. Oligonucleotides contained in such a library should form duplexes with their complements in a more narrow range of experimental conditions (temperature, salt concentration etc.) than the one characteristic for an oligonucleotide library with oligomers of the same length. Therefore, the hybridization experiments performed with isothermic libraries should result in a less number of experimental errors. This property of isothermic libraries prompted us to analyze their applicability in SBH. Their use should substantially limit a number of faults to be considered in the computational phase of the SBH approach.

It is easy to show that one isothermic library is not enough to perform an ideal hybridization experiment. For example, it is not possible to cover completely DNA chains composed of only C and G nucleotides by oligomers from a library of a temperature not divisible by 4. Moreover, a library of a temperature divisible by 4 is not sufficient to cover completely subsequences where one nucleotide of A (or T) type is surrounded by only G (or C) nucleotides. On the other hand, it is always possible to cover any DNA sequence by probes coming from two isothermic libraries of temperatures differing by 2 deg. Moreover, this coverage is such that in the sequence two consecutive oligonucleotides (from the libraries) have starting points shifted by at most one position.

From the above considerations it follows that in order to apply isothermic libraries to the SBH process, one should use two of them of temperatures differing by 2 deg. A construction of such libraries can be done in a linear number of steps [2]. Similarly like in case of standard SBH the combinatorial part of isothermic SBH with positive and negative errors is a strongly NP-hard problem. Integer linear programming formulation allows one for solving the problem by tabu search or branch and bound similarly to the approaches used in case of standard oligonucleotide libraries [3].

References

[1] Błażewicz, J., Formanowicz, P., Kasprzak, M., and Markiewicz, W.T., Method of sequencing of nucleic acids, The Patent Office of the Republic of Poland, Patent Application No P 335786, 1999.

[2] Błażewicz, J., Formanowicz, P., Kasprzak, M., and Markiewicz, W.T., Sequencing by hybridization with isothermic oligonucleotide libraries, Research Report RA-002/2000, Poznań Supercomputing and Networking Center, 2000.

[3] Błażewicz, J., Formanowicz, P., Kasprzak, M., Markiewicz, W.T., and Węglarz, J., DNA sequencing with positive and negative errors, *J. Comp. Biol.*, 6:113–123, 1999.

[4] Pevzner, P.A., l-tuple DNA sequencing: computer analysis, *J. Biomol. Struct. Dyn.*, 7:63–73, 1989.

[5] 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.

[6] Wallace, R.B., Johnson, M.J., Hirose, T., Miyake, T., Kawashima, E.H., and Itakura K., The use of synthetic oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA, *Nucleic Acids Res.*, 9:879–894, 1981.