BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Identification of regions in multiple sequence alignments thermodynamically suitable for targeting by consensus oligonucleotides: application to HIV genome

Olga V Matveeva[1], Brian T Foley[2], Vladimir A Nemtsov[3], Raymond F Gesteland[1], Senya Matsufuji[4], John F Atkins[1,5], Aleksey Y Ogurtsov[6] and Svetlana A Shabalina*[6]

Address: [1]Department of Human Genetics, University of Utah, Salt Lake City 84112-5330, USA, [2]Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545, USA, [3]MGGT, Ul. Lavochkina 23(A), 125502, Moscow, Russia, [4]Department of Biochemistry II The Jikei University School of Medicine 3-25-8 Nishi-Shinbashi, Minato-ku, Tokyo 105-8461, Japan, [5]Biosciences Institute, University College Cork, Ireland and [6]National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland 20814, USA

Email: Olga V Matveeva - olgam@howard.genetics.utah.edu; Brian T Foley - btf@atlas.lanl.gov; Vladimir A Nemtsov - vovanem@online.ru; Raymond F Gesteland - ray.gesteland@genetics.utah.edu; Senya Matsufuji - senya@jikei.ac.jp; John F Atkins - john.atkins@genetics.utah.edu; Aleksey Y Ogurtsov - ogurtsov@ncbi.nlm.nih.gov; Svetlana A Shabalina* - shabalin@ncbi.nlm.nih.gov

* Corresponding author

Published: 29 April 2004

BMC Bioinformatics 2004, **5**:44

Received: 14 February 2004
Accepted: 29 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/44

Abstract

Background: Computer programs for the generation of multiple sequence alignments such as "Clustal W" allow detection of regions that are most conserved among many sequence variants. However, even for regions that are equally conserved, their potential utility as hybridization targets varies. Mismatches in sequence variants are more disruptive in some duplexes than in others. Additionally, the propensity for self-interactions amongst oligonucleotides targeting conserved regions differs and the structure of target regions themselves can also influence hybridization efficiency. There is a need to develop software that will employ thermodynamic selection criteria for finding optimal hybridization targets in related sequences.

Results: A new scheme and new software for optimal detection of oligonucleotide hybridization targets common to families of aligned sequences is suggested and applied to aligned sequence variants of the complete HIV-1 genome. The scheme employs sequential filtering procedures with experimentally determined thermodynamic cut off points: 1) creation of a consensus sequence of RNA or DNA from aligned sequence variants with specification of the lengths of fragments to be used as oligonucleotide targets in the analyses; 2) selection of DNA oligonucleotides that have pairing potential, greater than a defined threshold, with all variants of aligned RNA sequences; 3) elimination of DNA oligonucleotides that have self-pairing potentials for intra- and inter-molecular interactions greater than defined thresholds. This scheme has been applied to the HIV-1 genome with experimentally determined thermodynamic cut off points. Theoretically optimal RNA target regions for consensus oligonucleotides were found. They can be further used for improvement of oligo-probe based HIV detection techniques.

Conclusions: A selection scheme with thermodynamic thresholds and software is presented in this study. The package can be used for any purpose where there is a need to design optimal consensus oligonucleotides capable of interacting efficiently with hybridization targets common to families of aligned RNA or DNA sequences. Our thermodynamic approach can be helpful in designing consensus oligonucleotides with consistently high affinity to target variants in evolutionary related genes or genomes.

Background

Finding optimal targets for oligonucleotides in multiple variants of related sequences is useful for a number of practical tasks. One of them is the design of oligonucleotide probes for RNA/DNA based pathogen detection assays. Beside PCR, such detection can be performed using strand displacement amplification (SD) [1,2], transcription-mediated amplification (TMA) [3], nucleic acid sequence-based amplification (NASBA) [4], hybridization protection assay [5], branched DNA signal amplification [6,7], *in situ* hybridization [8,9] or other techniques that are currently being developed and require oligonucleotides interacting with RNA or DNA as a basic step.

Sensitive detection of HIV RNA in plasma of infected persons is also achieved by methods that depend on binding of oligonucleotides to viral RNA sequences. Currently, RNA detection of some proportion of HIV-1 variants is not optimal, especially at low viral loads [10,11]. To develop more sensitive and far reaching detection assays, it is important to select HIV-1 RNA target regions where mutations are least disruptive for potential duplex formation with complementary oligonucleotides.

Computer programs for the generation of multiple sequence alignments such as "Clustal W" [12] allow detection of regions that are most conserved among many sequence variants. However, even for regions that are equally conserved, their potential utility as hybridization targets varies. Mismatches in sequence variants are more disruptive in some duplexes than in others. Additionally, the propensity for self-interactions amongst oligonucleotides targeting conserved regions differs and the structure of target regions themselves can also influence hybridization efficiency.

The currently existing methods that predict effective oligonucleotide primers for performing PCR from DNA templates work well for those applications where relatively stringent conditions are employed. This is because PCR experimental design greatly simplifies the prediction problem: hybridization is performed at relatively low ionic strength and high temperature. Under these conditions, oligonucleotide and target secondary structures are relatively unimportant. The secondary structure of an oligonucleotide or its target RNA may be not a significant problem for influencing hybridization efficiency at 65°C or higher temperatures, but it becomes a problem at temperatures close to 37°C. These temperatures are frequently used for oligo-RNA hybridization in a number of different RNA detection assays.

This work suggests thermodynamic filtering procedures to select optimal consensus oligonucleotide targets in multiple sequence variants that can be used for RNA detection assays performed at 37°C.

Results and discussion

The scheme developed for discrimination of conserved regions in multiple sequence variants is based on their potential to serve as efficient hybridization targets for oligonucleotides and involves several steps: First, creation of a consensus sequence of RNA or DNA from aligned sequence variants with specification of the lengths of fragments to be used as oligonucleotides in the analyses. Second, selection of DNA oligonucleotides that have pairing potential, greater than a defined threshold, within the set or subset of the aligned RNA sequences. Third, elimination of DNA oligonucleotides that have self-pairing potentials for intra- and inter-molecular interactions greater than defined thresholds. The consensus RNA subsequences complementary to the remaining set of oligonucleotides are preferred potential targets for hybridization.

The discrimination scheme described above was applied to the HIV genome where the need to identify hybridization targets is obvious. Here we present the result of analysis for the HIV-1 *gag* gene, data for the complete HIV genome are available at: http://www.gesteland.genetics.utah.edu/HIV/37C_targets.html. For each successive fragment of consensus sequence varying in size from 20 to 35 nt., oligonucleotides forming stable duplexes with 90% of RNA variants (free energies (ΔG°_{37}) ≤ -30 kcal/mol) and little self-structure with (ΔG°_{37}) ≥ -8 kcal/mol for inter-oligonucleotide pairing and (ΔG°_{37}) ≥ -1.1 kcal/mol for intra-molecular pairing, were selected [13,14].

Theoretically optimal hybridization targets are shown in Figure 1. Figure 2 shows that only a sub-set of conserved target fragments in the *gag* gene is "optimal" for hybridization with oligonucleotides at 37°C. This histogram demonstrates that the conservation values for 30 nucleotide *gag* windows vary from 68% to 95%.

The *gag* consensus sequence yields a total number of 23704 complementary oligonucleotides ranging in size from 20 to 35 mers. A set of 1747 oligonucleotides that is 14 times smaller than the initial one, remains after application of the thermodynamic discrimination steps described here. Limited work has been performed on simultaneous combinations of thermodynamic or homology analyses for predicting optimal universal targets in related RNA sequences for oligonucleotide hybridization [15-19]. The present work employs a distinctive scheme and shows its utility. The key ingredients in the present scheme are experimentally derived thermodynamic discriminatory steps. Decisions about the suitability of a particular target region are determined by a set of

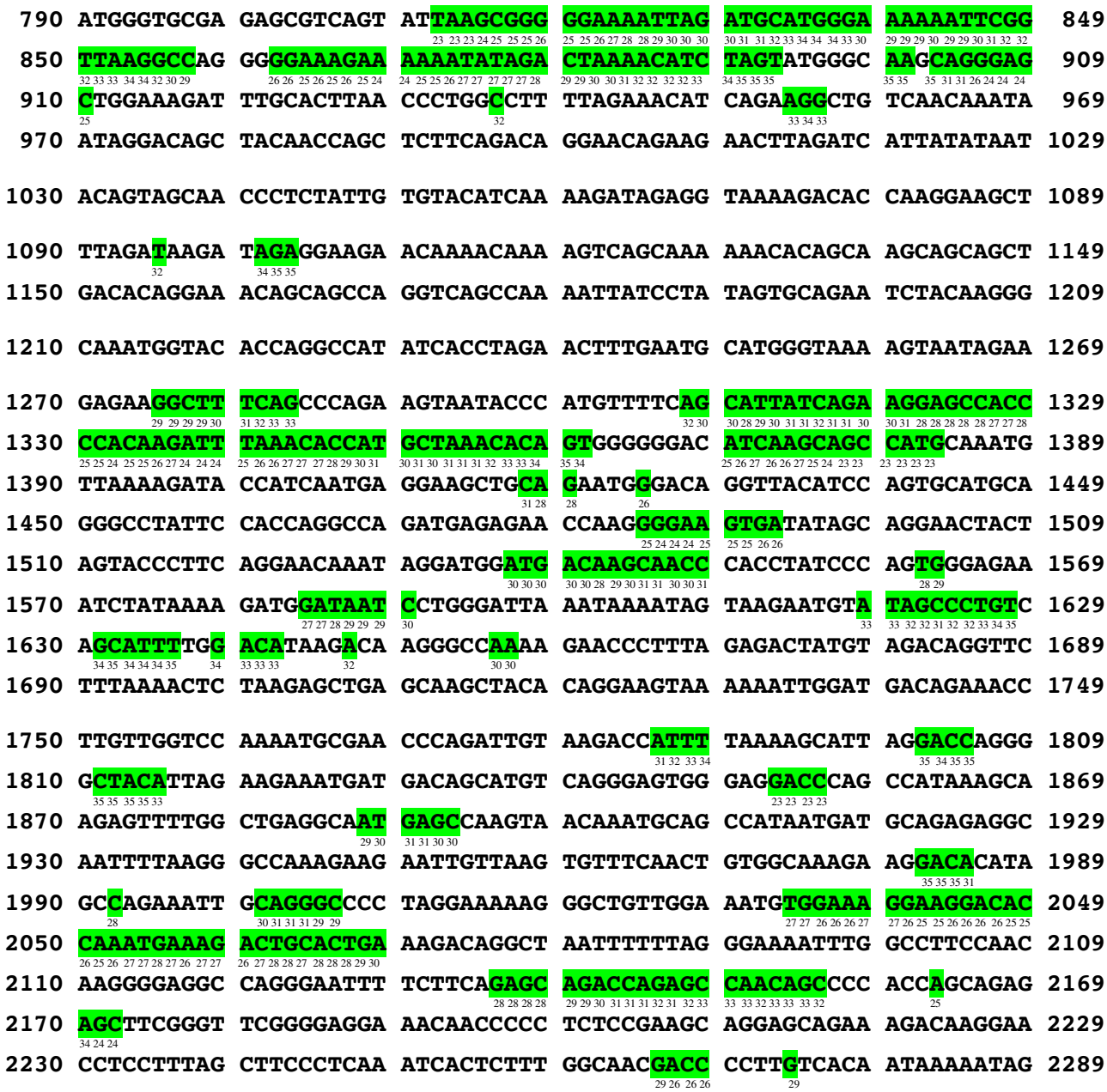

Figure 1

Gag consensus sequence. Last nucleotides in the theoretically optimal target regions are highlighted. The range of fragments that were analyzed was from 23 to 35-mers. The length of optimal region is shown below the highlighted nucleotide. Only numbers for shortest regions in the sets that correspond to each highlighted nucleotide are shown.

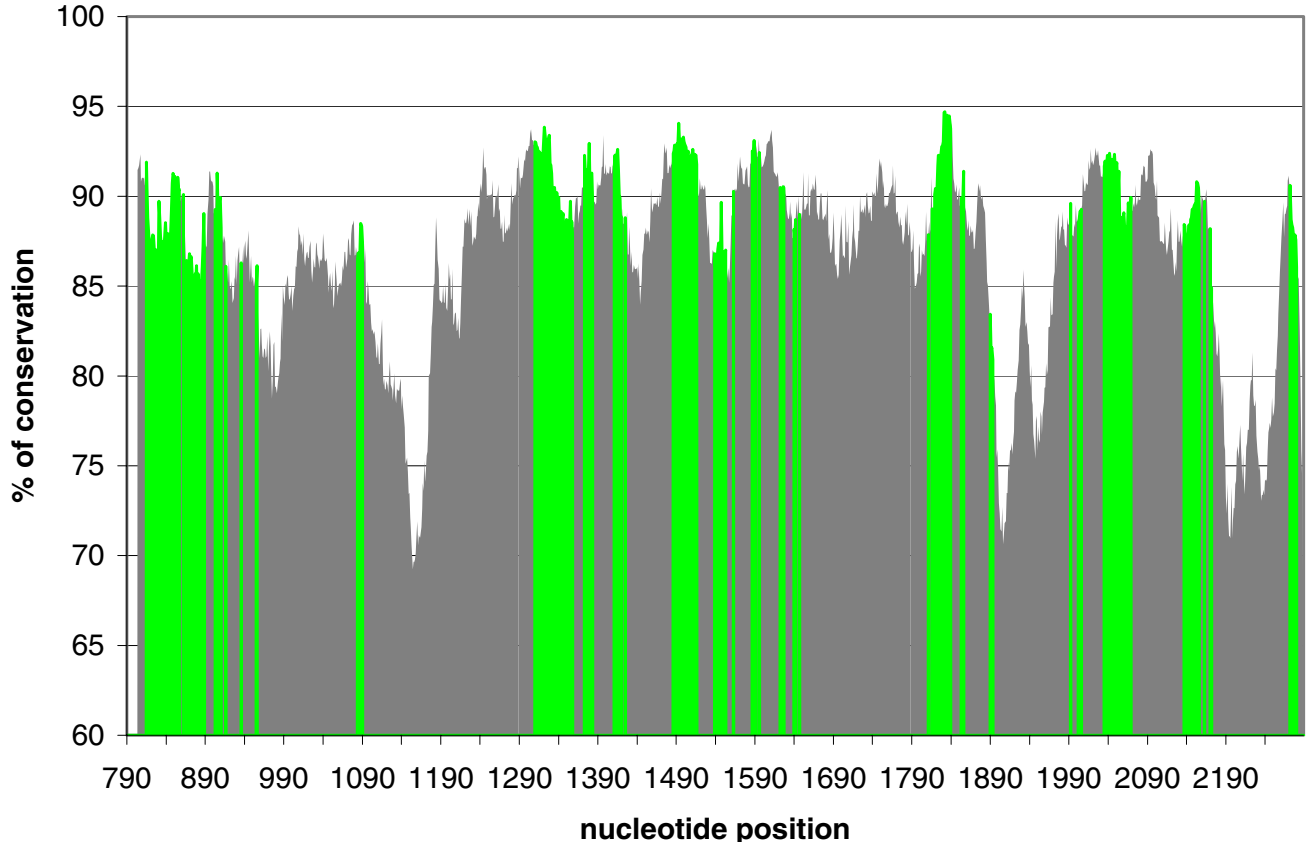

Figure 2
Gag plot of conservation made with window of 30 nucleotides. Average conservation for each consequent 30 nucleotides is shown. Conserved regions that are thermodynamically optimal for oligonucleotide targeting are highlighted in green. The last nucleotide of each fragment is highlighted in conservation histogram.

thresholds, which were found after analysis of the efficiency of oligonucleotides in the experiments performed *in vivo* and in *vitro* [13,14]. Oligonucleotides that form stable duplexes with RNA (free energies (ΔG°_{37}) ≤ -30 kcal/mol) and little self-structure are statistically much more likely to be active than molecules, which form less stable oligonucleotide-RNA hybrids or more stable self-structures. To achieve optimal statistical preference, the values for self-interaction should be (ΔG°_{37}) ≥ -8 kcal/mol for inter-oligonucleotide pairing and (ΔG°_{37}) ≥ -1.1 kcal/mol for intra-molecular pairing [13,14]. Selection of oligonucleotides with these thermodynamic values in the analyzed experiments would have increased the proportion of active oligonucleotides by as much as six fold.

Oligonucleotide length correlates with the numbers of theoretically optimal RNA targets obtained after conservation and thermodynamic selection procedures. More optimal targets can be detected for longer oligonucleotides (Figure 3).

The temperature used for the experiments from which the thermodynamic thresholds were derived, is 37°C. Application of these thresholds in the current work yields hybridization target regions that are optimal for the same temperature. The list of selected regions for oligonucleotide hybridization targeting is good for procedures that involve oligonucleotide RNA pairing at 37°C such as branch DNA detection technology and often reverse transcription. For PCR that requires higher temperature, other thermodynamic thresholds obviously need to be used.

Future improvement of the predictions requires incorporation of target RNA secondary structure considerations. RNA secondary structure based thermodynamic filtration

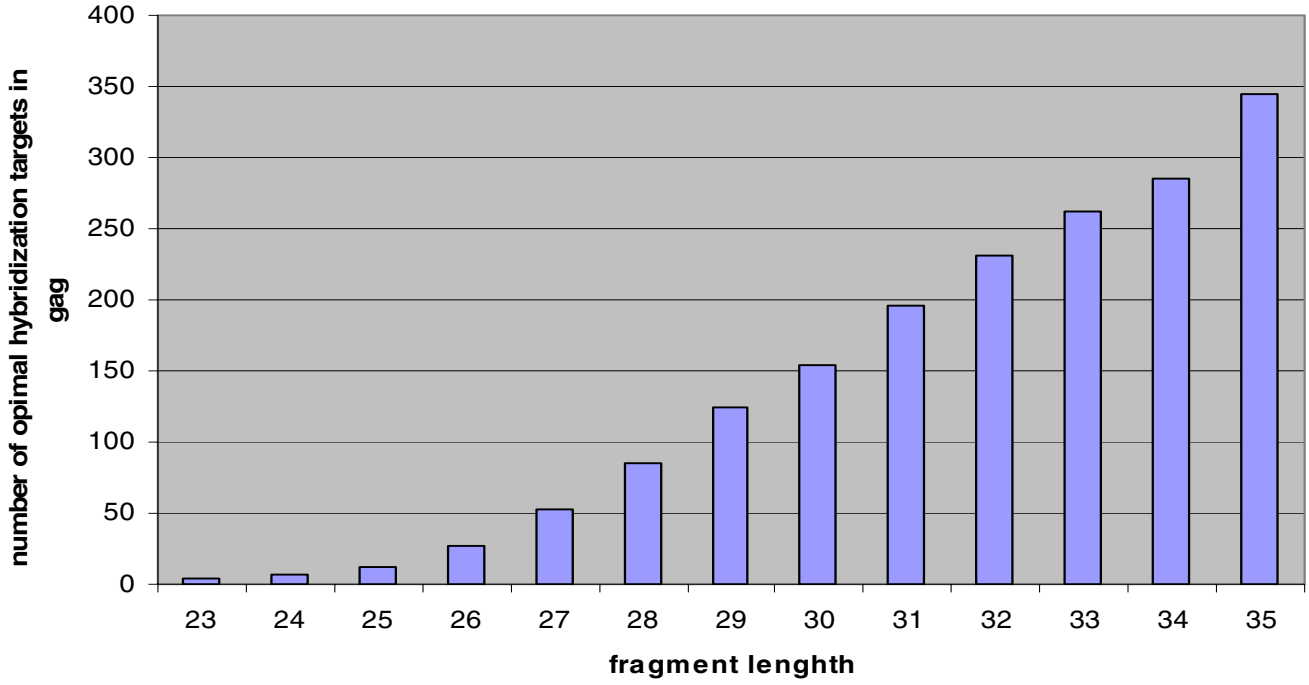

Figure 3
The number of theoretically optimal RNA targets obtained with each possible length of oligonucleotide, in the range from 23 to 35-mers.

can be added when information about optimal thresholds for discrimination becomes available from analyses of experimental databases. A further improvement needed is combination of the software elements for different steps of discrimination analysis into one package with a common input and output. This can make the suggested analysis faster and more convenient for a broad range of users.

Whether or not an oligo-molecule will be a good hybridization probe for HIV detection also depends on the specificity of the oligo-RNA interactions. It is possible for an oligo-probe to be in the optimal range of thermodynamic values described above, but still to cross-hybridize with unintended RNA targets (human mRNAs for example). The BLAST analysis tool may be used to discriminate oligo-probes with high potential to cross-hybridize with non-specific targets [20]. We applied the BLAST program for this purpose and ranked oligos that passed all the thermodynamic filters in accordance with frequencies of non-specific targets in the human and HIV genomes (see ftp://ftp.ncbi.nih.gov/pub/kondrashov/HIV). To detect a maximum of sequence similarities the word size for the BLAST search was set to the smallest allowed value (-W 7). However, BLAST is unable to generate thermodynamic values that could be linked to the experimental oligo-probe hybridization, and the BLAST score alone does not allow prediction of oligoprobe hybridization specificity and sensitivity.

The consensus oligonucleotides for targets that were selected after rounds of discrimination analysis should be prime candidates for sensitive viral detection procedures or experiments that require efficient oligonucleotide-RNA interaction for the broad range of viral variants. The set of oligonucleotides for *gag* that remains after homology and thermodynamic selection is 14 times smaller than the initial set of all possible oligonucleotides in this range. Statistical and thermodynamic analyses performed with experimental oligo-probe datasets [13,14] suggests that approximately 70% of the oligonucleotides from this theoretically selected set will demonstrate consistency in hybridization behavior with different HIV-1 representatives of group M viruses. It is likely that the suggested thermodynamic selection rules for finding the most efficient hybridization oligo-probes can also be applied to siRNA design. However, the thermodynamic

threshold values and position scores of matches to target sequences have to be adjusted for this purpose, and is the subject of separate work (in preparation).

Conclusions

A selection scheme with thermodynamic thresholds and new software were developed in this study. The selection scheme can be used for any purpose where there is a need to design optimal consensus oligonucleotides capable to interact efficiently with hybridization targets common to families of aligned RNA or DNA sequences. It employs creation of a consensus sequence of RNA or DNA from aligned sequence variants and filtering procedures with experimentally determined thermodynamic cut off thresholds. This scheme has been applied to the HIV-1 genome and theoretically optimal RNA target regions for consensus oligonucleotides were found. They can be further used for improvement of oligo-probe based HIV detection techniques.

Methods

Consensus sequence and multiple sequence alignments

The consensus sequence for HIV-1 variants (group M) and multiple sequence alignments created by Los Alamos Laboratory staff [21] were used in this work.

Plot of conservation

The average percentage of conservation of each 30 consecutive nucleotides in multiple sequence alignments (based on division of the sum of percentage conservation of each nucleotide by the number of nucleotides) was calculated using the program created for this study.

Evaluation of the potential for intra-molecular and inter-molecular self-interaction of DNA oligonucleotides

Calculations of thermodynamic properties of oligonucleotides were done with the help of OligoScreen program from RNAStucture 3.71 package [22].

Evaluation of pairing potentials among DNA oligonucleotides and target RNA variants

A computer program AlignScan was created to evaluate, by ΔG°_{37} calculations, the pairing potential of each DNA or RNA consensus fragment with divergent RNA variants. The program is available for downloading from the site http://gesteland.genetics.utah.edu/HIV/AlignScan.zip.

AlignScan requires aligned sequence variants as an input file (FASTA format). A sample input file gag.cgi with the list of aligned sequence variants of HIV-1 *gag* genes is included for downloading. AlignScan also requires for input the temperature, the range of lengths of oligoprobes, ΔG°_T threshold value and percentage of sequences for which this threshold have to be valid. AlignScan creates consensus sequence from aligned sequence variants and ΔG°_T values are calculated for all complementary duplexes between each successive fragment of consensus sequence and the corresponding fragment in all sequence variants. The AlignScan output file is in txt format and includes all oligonucleotides of inputted length from the consensus sequence with ΔG°_T values for duplexes between each consensus oligonucleotide and the corresponding complementary target sequence variants below the threshold. The AlignScan output file can be further used as input file for the OligoScreen program. The output file can be also opened for analysis by other programs, such as Microsoft Excel.

List of abbreviations

PCR – polymerase change reaction

SD – strand displacement amplification

TMA – transcription mediated amplification

NASBA – nucleic acid sequence-based amplification

HIV-1 – Human Immunodeficit Virus type 1

Authors' contributions

O.V. Designed the study, coordination and drafted the ms

B.T. Helped to conceive the idea of the study

V.A. Created the program AlignScan

A.Y. Helped to create, test and describe the program AlignScan

R.G. Helped in study design

S.M. Participated in sequence analysis

J.A. Helped in study design and ms writing

S.A. Performed computer analyses and designed the study

Acknowledgements

We thank Alexander Tsodikov for helpful discussions and comments. Mr Andy Hammer for updating and management of our downloadable files. We also thank Nikolay Spiridonov for critical reading of the manuscript. O.M. and J.F.A. were supported by a sub-contract from the Japan Society for the promotion of Science, and J.F.A. was also supported by NIH grant GM48152 and by Science Foundation Ireland.

References

1. Walker GT, Fraiser MS, Schram JL, Little MC, Nadeau JG, Malinowski DP: **Strand displacement amplification – an isothermal, in vitro DNA amplification technique.** *Nucleic Acids Res* 1992, **20:**1691-1696.
2. Walker GT, Little MC, Nadeau JG, Shank DD: **Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system.** *Proc Natl Acad Sci USA* 1992, **89:**392-396.

3. Kacian DL, Fultz TJ: **Nucleic Acid Sequence Amplification Methods.** *U.S. Patent No. 5.399.491* 1995.
4. Compton J: **Nucleic acid sequence-based amplification.** *Nature* 1991, **350:**91-92.
5. Arnold LJ Jr, Hammond PW, Wiese WA, Nelson NC: **Assay formats involving acridinium-ester-labeled DNA probes.** *Clin Chem* 1989, **35:**1588-1594.
6. Urdea MS, Wilber JC, Yeghiazarian T, Todd JA, Kern DG, Fong SJ, Besemer D, Hoo B, Sheridan PJ, Kokka R, Neuwald P, Pachl CA: **Direct and quantitative detection of HIV-1 RNA in human plasma with a branched DNA signal amplification assay.** *Aids* 1993, **7(Suppl 2):**S11-14.
7. Urdea MS: **Branched DNA signal amplification Biotechnology.** *(N Y)* 1994, **12:**926-928.
8. DeLong EF, Wickham GS, Pace NR: **Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells.** *Science* 1989, **243:**1360-1363.
9. Amann RI, Ludwig W, Schleifer KH: **Phylogenetic identification and in situ detection of individual microbial cells without cultivation.** *Microbiol Rev* 1995, **59:**143-169.
10. Chew CB, Zheng F, Byth K, Van Asten M, Workman C, Dwyer DE: **Comparison of three commercial assays for the quantification of plasma HIV-1 RNA from individuals with low viral loads.** *Aids* 1999, **13:**1977-1978.
11. Debyser Z, Van Wijngaerden E, Van Laethem K, Beuselinck K, Reynders M, De Clercq E, Desmyter J, Vandamme AM: **Failure to quantify viral load with two of the three commercial methods in a pregnant woman harboring an HIV type 1 subtype G strain.** *AIDS Res Hum Retroviruses* 1998, **14:**453-459.
12. Higgins DG, Sharp PM: **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.** *Gene* 1988, **73:**237-244 [http://www.ebi.ac.uk/clustalw].
13. Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, Atkins JF: **Thermodynamic calculations and statistical correlations for oligo-probes design.** *Nucleic Acids Res* 2003, **31:**4211-4217.
14. Matveeva OV, Mathews DH, Tsodikov AD, Shabalina SA, Gesteland RF, Atkins JF, Freier SM: **Thermodynamic criteria for high hit rate antisense oligonucleotide design.** *Nucleic Acids Res* 2003, **31:**4989-4994.
15. Lucas K, Busch M, Mossinger S, Thompson JA: **An improved microcomputer program for finding gene- or gene family-specific oligonucleotides suitable as primers for polymerase chain reactions or as probes.** *Comput Appl Biosci* 1991, **7:**525-529.
16. Dopazo J, Rodriguez A, Saiz JC, Sobrino F: **Design of primers for PCR amplification of highly variable genomes.** *Comput Appl Biosci* 1993, **9:**123-125.
17. Proutski V, Holmes EC: **Primer Master: a new program for the design and analysis of PCR primers.** *Comput Appl Biosci* 1996, **12:**253-255.
18. Kel A, Ptitsyn A, Babenko V, Meier-Ewert S, Lehrach H: **A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the G protein-coupled receptor protein superfamily.** *Bioinformatics* 1998, **14:**259-270.
19. Gibbs A, Armstrong J, Mackenzie AM, Weiller GF: **The GPRIME package: computer programs for identifying the best regions of aligned genes to target in nucleic acid hybridisation-based diagnostic tests, and their use with plant viruses.** *J Virol Methods* 1998, **74:**67-76.
20. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.** *Nucleic Acids Res* 2003, **31:**3057-3062.
21. Gaschen B, Kuiken C, Korber B, Foley B: **Retrieval and on-the-fly alignment of sequence fragments from the HIV database.** *Bioinformatics* 2001, **17:**415-418. http://hiv-web.lanl.gov/content/hiv-db/CONSENSUS/M_GROUP/Consensus.html, http://hiv-web.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html
22. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH: **Predicting oligonucleotide affinity to nucleic acid targets.** *RNA* 1999, **5:**1458-1469 [http://128.151.176.70/RNAstructure.html].