

The microarray manual curation tool (MMCT): A Web server for microarray probe evaluations

Dan Tulpan*, Luc Belliveau, Serge Léger

Institute of Information Technology National Research Council of Canada Moncton, New Brunswick, E1A 7R1, Canada; Phone: +1 (506) 861-0958 Fax: +1 (506) 851-3630; Dan Tulpan - E-mail: dan.tulpan@nrc-cnrc.gc.ca. *Corresponding author.

Received January 7 2010, Accepted February 8 2010, Published February 28 2010

Abstract:

Quality control of probe sequences is a major concern in microarray technology. The presence of poor quality probes has a negative impact on the microarray data analysis process. The Microarray Manual Curation Tool (MMCT) is a web server application that provides computational and visual means to investigate the quality of individual probes for oligo microarrays. The MMCT quality metrics assess the free energy of hybridization and the secondary structure of duplexes formed by selected targets and probes, which are specific to various microarray platforms.

Availability: <http://www.nrcbioinformatics.ca/mmct>

Keywords: Bioinformatics, web server, microarrays, visualization.

Background:

DNA microarrays [1, 2, 3, 4] are widely used tools capable of measuring expression levels of genes at genomic scale. Microarrays heavily rely on the process of hybridization between DNA sequences affixed to the surface of the microarray (probes) and DNA or RNA sequences present in solution (targets). The process of hybridization between a probe and a target relies on complementary A-T and C-G base pairs. The gene expression level is thus estimated by optical measurements of the fluorescence intensity obtained from the interaction of labeled targets and surface-bound probes. It is widely believed that gene expression fluorescence intensities are typically given by hybridizations of probes and their corresponding targets, but often, strong non-complementary hybridization events (cross-hybridizations) are observed. The existence of cross-hybridization is attributed to many factors, the most prominent being poor probe designs. Some freely available tools allow statistical quality assessment for specific microarray platforms, such as Affymetrix [5], Illumina [6], cDNA arrays [7]. Other tools have the computational capabilities to detect imperfections in spotting [8], hybridization [9] or intensity values [10], but most of them do not provide interactive means of visualization.

This paper introduces MMCT (the Microarray Manual Curation Tool) – a web server application that provides computational and visual insights into microarray probe and target sequence hybridizations. The tool provides the means to explore the quality of individual probes on virtually any oligo microarray chipset (currently only Affymetrix data is loaded) by performing computational hybridization simulations for each probe and consecutive sub-sequences of user-specified length taken along a pre-specified target.

Methodology:

The core of MMCT is a MySQL database containing microarray sequence information relevant to Human chips from Affymetrix website [11]. Wrapped around the database, we built a web interface powered by PHP, Java-Script, Python and Ajax. The computational aspect of the web application consists of a modified version of the free energy prediction tool PairFold [12], and a set of Perl pre-processing and GD-powered drawing scripts.

Input:

MMCT takes as input the microarray platform, the probe (and target) identifiers (Affymetrix probe set IDs) or directly the user input sequences, together with the length (L) of a subsequence of the target and computes the hybridization free energies for all length L subsequences of the target and the probe. The free energies are

computed using PairFold - a publicly available tool that predicts minimum free energy secondary structures formed by two input DNA or RNA molecules.

Computations:

When a chip platform together with probe and target identifiers are passed as input to the web interface, MMCT queries the MySQL database containing up-to-date information about the chip and extracts the corresponding probe and target sequences. MMCT provides also an auto-completion Ajax based mechanism that helps the user to locate and select the desired probe and target faster. The extracted probe and target sequence are passed as input to a Perl script that calls PairFold [2] on consecutive sub-sequences of the target and free energy estimates of hybridization strength are calculated.

The hybridization strength of two DNA sequences is typically modeled using the Nearest-Neighbor Thermodynamic model originally developed in Crothers Lab [13] and is estimated computationally using minimum free energies. The lower the free energy is, the stronger the hybridization of a probe and a target subsequence is. Figures 1 and 2 show the free energy profiles of a perfect-match and, respectively, an imperfect-match probe from probe set 1213_at on Affymetrix chipset HG_U95Av2.

Output:

Once the input is provided and the free energies are estimated using predefined thermodynamic parameters, a plot with free energy profiles is displayed. The plot shows two curves: (i) a blue curve representing probe versus sub-target hybridization free energies and (ii) a red curve representing sub-target perfect match hybridization free energies (see Figures 1 and 2). The red curve is used as a baseline for identification of potential matches along a target. If the probe perfectly matches a subsequence of the target, then the red and the blue curves intersect. The vertical distance between red and blue points with equal x-axis coordinates decreases if the probe hybridizes stronger with the sub-target sequence. The bi-dimensional plot represents the start positions of the sub-targets on the x-axis and the estimated free energies measured in kcal/mol on the y-axis. MMCT provides also means for visualizing secondary structures of probe-target interactions.

Figure 3 shows the secondary structure formed by probe interrogating position 3582 from probe set 1213_at on Affymetrix chipset HG_U95Av2 and a subsequence of the corresponding target. Intuitive browsing features allow the user to explore the relationships between the probe and target sequences and secondary structure information via

zooming in and out and automatic highlighting of structural nucleotides when the text-input cursor is moved along one of the sequences. The probe sequence is depicted in black and has gray 5' and 3' ends, while the sub-target sequence is depicted in blue with

light blue 5' and 3' ends. All plots produced by MMCT can be exported as Portable Network Graphics (PNG) images.

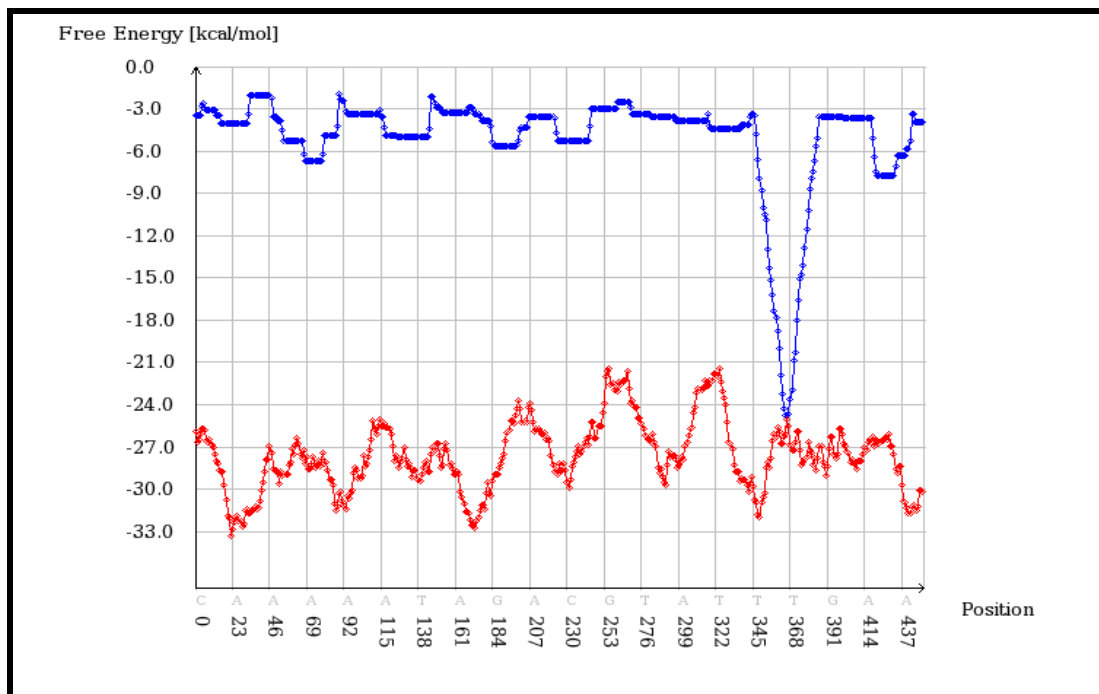


Figure 1: Free energy profile for the probe interrogating position 3582 from probe set 1213_at on Affymetrix chipset HG_U95Av2.

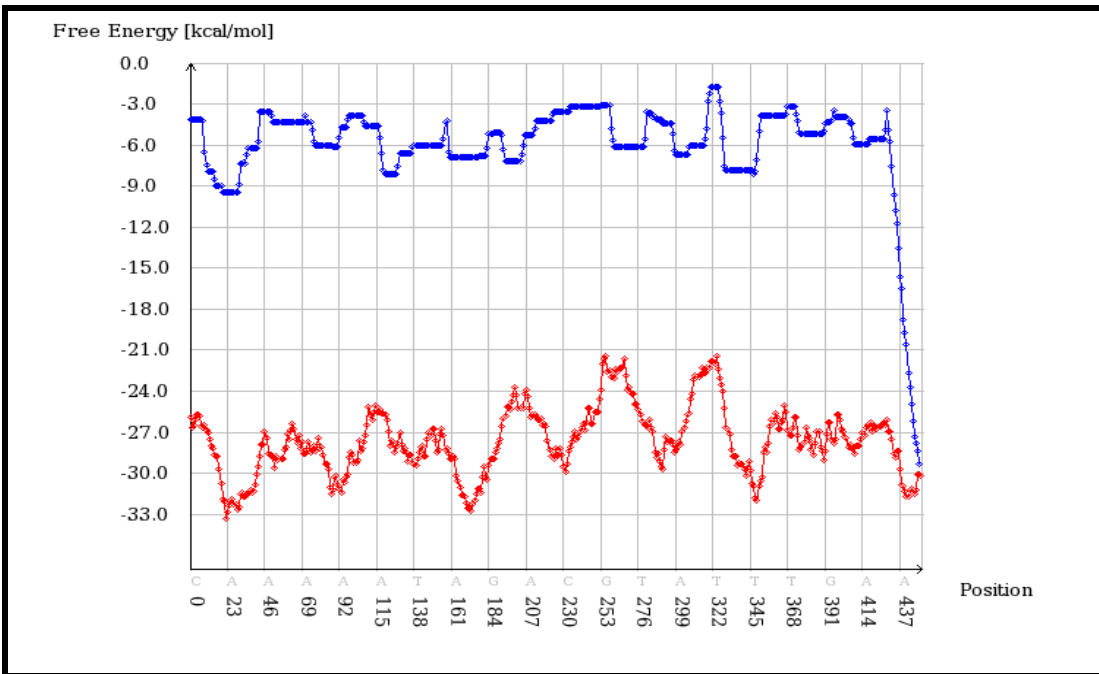


Figure 2: Free energy profile for the probe interrogating position 3666 from probe set 1213_at on Affymetrix chipset HG_U95Av2.

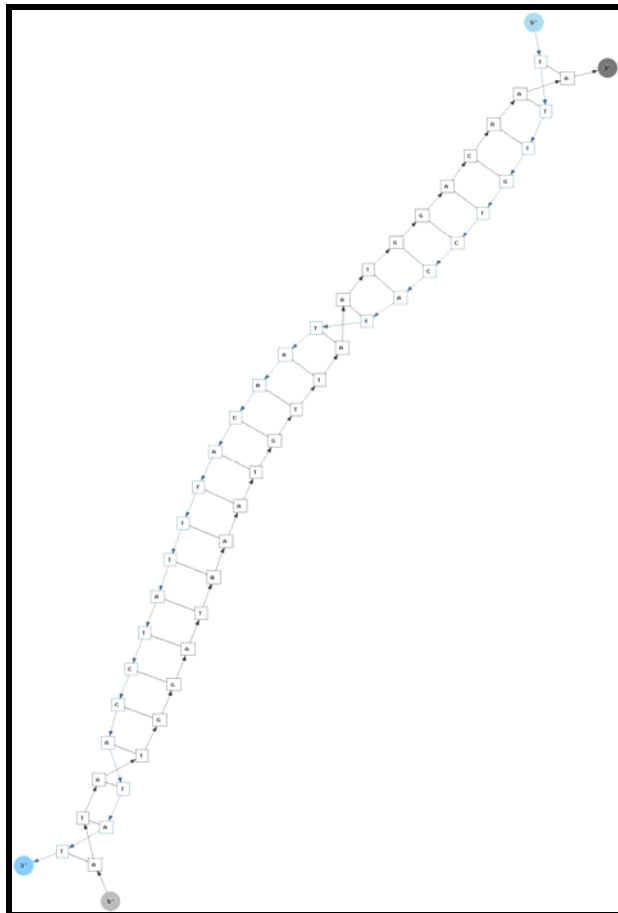


Figure 3: Secondary structure for the probe interrogating position 3582 from probe set 1213_at on Affymetrix chipset HG_U95Av2 and a subsequence of the corresponding target.

Future development:

We continue the development of this tool by integrating alternative hybridization model parameters and by adding more attractive data browsing features including annotation information relevant to each microarray probe.

Acknowledgments:

Funding for this research was kindly provided by the National Research Council of Canada.

References:

- [1] PO Brown & D Botstein, *Nature Genet.*, (1999) **21**: 33 [PMID: 9915498]
- [2] RJ Lipshutz *et al.*, *Nature Genetics*, (1999) **21**: 20 [PMID: 9915496]
- [3] DD Lockhart, *Nat Biotechnol.*, (1996) **14**: 1675 [PMID: 9634850]
- [4] MS Schena, *Science*, (1995) **270**: 467 [PMID: 7569999]

- [5] C Parman, & C Halling, *AffyQCReport: QC Report Generation for affyBatch objects*, (2005) R package version 1.17.0
- [6] MJ Dunning, *et al.*, *Bioinformatics*, (2007) **23**: 2183 [PMID: 17586828]
- [7] A Buness *et al.*, *Bioinformatics*, (2005) **21**: 554 [PMID: 15454413]
- [8] Q Li *et al.*, *Bioinformatics*, (2005) **21**: 2875 [PMID: 15845656]
- [9] Petri,A. *et al.*, *BMC Bioinformatics*, (2004) **5**: 12 [PMID: 15018654]
- [10] A Kauffmann *et al.*, *Bioinformatics*, (2009) **25**: 415 [PMID: 19106121]
- [11] <http://www.affymetrix.com/support/index.affx/>
- [12] M Andronescu *et al.*, *Nucleic Acids Research*, (2003) **31**: 3416 [PMID: 12824338]
- [13] DM Crothers & BH Zimm, *J. Mol. Biol.*, (1964) **9**: 1 [PMID: 14200383]

Edited by P Kanguane

Citation: Tulpan *et al.* Bioinformatics 4(8): 344-346 (2010)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.