# “DEQOR: a web-based tool for the Design and Quality Control of siRNAs”

Andreas Henschel, Frank Buchholz and Bianca Habermann

## Supplementary Material

**Figure S1: Parameter Testing for DEQOR**. Each penalty parameter used in a DEQOR analysis was tested in a high-throughput manner using 3500 randomly selected mRNAs from the human UniGene dataset. (A) Average quality scores plotted against used penalty score for poly-nucleotide (poly-nt) stretches. While poly-A and poly-T penalties showed a steeper increase, the increase of poly-G and poly-C stretches displayed a rather flat behavior. The penalty for each poly-nt stretch was increased with a fixed penalty of 7 for the remaining nucleotides. (B) Penalties for reverse asymmetry or symmetry of siRNAs. Symmetric behavior of siRNAs increased rather steeply with increasing penalties, while reverse asymmetry showed a moderate increase with increasing penalties. (C) Different penalty parameters were tested for the GC – content of siRNAs. 4 functions were tested, whereby 2 functions preferentially selected siRNAs with 50% GC-content (Fct2 and Fct4), Fct3 assumed that a low GC – content (between 0% and 30%) would be optimal and, finally, Fct1 assumed that a GC – content between 20% and 50% would yield optimal silencing potential and was therefore selected for the program DEQOR. Penalty functions are as follows:

Fct1: f1(x)=abs(x-50);

Fct2: f2(x)=abs(x-50)/2;

Fct3: f3(x)= 0 for $x \leq 20$
(x-20)/3 for $x > 20$

Fct4: f4(x)= 20-x for $x \leq 2$
0 for $20 < x < 50$
(x-50)/2 for $x \geq 50$

(D) The distribution of the % GC – content for all siRNAs resulting from selected input sequences. The GC – content was normally distributed, with a peak between 45% and 50%.

**Figure S2: Parallelization of DEQOR**. The high number of BLAST-searches that have to be carried out by DEQOR is a considerable load for any standard computer. Essentially, the number of BLAST-jobs that have to be carried out for a single input query equals the length of the query minus the length of a single siRNA frame plus 1. For a query of 500 nt length and a window size of 21 nt, 480 BLAST-jobs are required. DEQOR analysis on a single Linux box is therefore a computationally time-intensive task. One DEQOR analysis against the human transcriptome with a sequence of an average length (500 bp) takes about 63 seconds on a two-processor Linux machine. This is a major draw back for siRNA design and control, especially when large data sets have to be analyzed, for instance in a high-throughput screen. To speed up running time of DEQOR, the program was parallelized on a Linux-cluster. The first step of a DEQOR analysis (recognition of the input sequence by a full-length *blastN*-search) is carried out on the master node. After that, the sequence is fragmented into pieces of 50 nt, representing one job package. In order not to lose sequence information, an overlap between sequence fragments is considered, which depends on the window size chosen (in case 21 nt is selected for fragmenting the input sequence, the overlap would for instance be 20 nt). In the next step, the job

packages are sent to the nodes in the cluster, which carry out the *in silico* digestion, quality control and the *blastN* searches of each siRNA against the selected transcriptome/genome at the local node. After finishing the job package, the node returns the results to the master and receives a new job package, therefore handling one job package at a time. After having received the last package from the nodes, the final step of analysis is again carried out on the master node, which sorts the siRNAs according to the score and produces the graphical– and html– output. Since final evaluation and network trafficking is minimal, parallelization of DEQOR leads to a gain in speed that is proportional to the number of nodes in the cluster. As an example, a DEQOR analysis of a 504 bp sequence against the human transcriptome required 63 seconds on a single, 2-processor node, while the same job could be processed in 21 seconds on 5 2-processor nodes, which is a gain in speed of approximately a factor of 3.

**Figure S3: Experimental validation of DEQOR parameter settings.** DEQOR output for three mammalian genes knocked down previously using esiRNA (1). While Clathrin Light Chain (LC) (A) and CDK2 (B) show large regions of high quality siRNAs, c-myc only contains approximately 50 nt of a high quality silencing region (C). LC and CDK2 furthermore contain over 30% high-quality siRNAs, while c-myc only contains approximately 15%. These data are correlated with the silencing quality of the individual probes that was determined experimentally. The accession numbers of used genes are: LC: M20471 (complete CDS); CDK2: X61622 (complete CDS); c-myc: V00568 (nt 566-1245).

## References

1. Yang, D., Buchholz, F., Huang, Z., Goga, A., Chen, C.Y., Brodsky, F.M. and Bishop, J.M. (2002) *Proc Natl Acad Sci U S A*, **99,** 9942-9947.