siRNA specificity searching incorporating mismatch tolerance data. – Supplementary Data

Alistair M. Chalk, Erik L.L. Sonnhammer.

We here review the sequence requirements for non-specific off-target siRNA effects. Using specificity data available from different sources, we create scoring schemes for assessing the likelihood of a non-specific effect occurring. These scoring schemes are implemented in a web server that allows the user to check the specificity of a given siRNA using a combination of methods. The server finds potential off-target matches and ranks them according to a scoring system we introduce. We then apply the specificity search to a dataset of siRNAs with known efficacy derived from the siRNAdb (Chalk et al., 2005; http://sirna.sbc.su.se), which incorporates data from the HuSiDa (Truss et al., 2005) database. An analysis of the potential non-specific effects of these siRNAs is reported. The results of this analysis have been integrated into siRNAdb. We use the following numbering convention for siRNA nucleotide positions: position 1 refers to the 5' most base of the siRNA target (guide strand) and position 19 refers to the 3' most base.

Method and Results

Specificity scoring schemes

We developed a specificity score (SS) based on the data from Du *et al.* (Du et al., 2005) to allow the evaluation of the affinity between an siRNA and any potential target. The score, denoted as $SS = 100 - \Sigma W_i \cdot M_i$. Where W = weighting for position i, M = 0 if match, 1 if mismatch at position i. The weighting can be defined as avg (W_{AVG}) or max (W_{MAX}), the average efficacy and the maximum efficacy of the mismatches respectively. The weightings are shown in figure S1, indicating a clear difference in tolerance between mismatches in the terminal bases and those in the centre of the siRNA, particularly positions 5-11. Using this scoring scheme, the output reflects the difference in the importance of mismatches at different positions on the target sequence. Of the two weightings presented, SS_{MAX} will give a higher penalty for mismatches. SS_{AVG}, the more tolerant weighting is used in further analysis. A modified version of the scoring scheme (SS2) incorporates the fact that mismatches between particular nucleotides are less critical. In this scheme, M is set to zero if a match is found, 0.8 if a mismatch forms a wobble base pair, and 1 for other mismatches. The specificity score can be used as a rough estimate of the level of silencing that can be expected for off-target genes. However, since we are combining silencing effects that were measured as single site changes, the score should not be seen as a quantitative predictor of gene silencing.

In addition we examined the off-target genes found by Jackson *et al.* (Jackson et al., 2003) to be consistently affected by a given siRNA and developed a set of criteria for off-target effects. These criteria are as follows: a) Matches with 14 or more contiguous matches (in range 3-17), b) 8+ contiguous matches at sense 3' end (range 11-19) and c) Positions 12 and 13 are 100% conserved. These rules account for 7/9 of the observed non-specific hits. Criterion a) represents a conservative view of specificity requirements, while criterion b) is more promiscuous and finds a large number of false positives (discussed below). The importance of criterion c) is uncertain and should be approached with a measure of skepticism.

Analysis of microarray based specificity data

In order to estimate the false positive rate of using the microarray derived criteria, we performed a WU-BLAST search using the highest possible sensitivity against the human RefSeq database (as an estimate of the chip contents). This resulted in a total of 2019 hits or varying identity to the query siRNA. We proceeded to filter these hits based on criteria derived from Jackson *et al*. We found one additional gene (criteria a) that was either unaffected by the siRNA, or not present on the chip. If the gene was present and unaffected then ~ 4/5 hits of the same similarity level have a non-specific effect. In contrast, 62 criterion (b) hits were found in the database. The observation of this high number leads us to conclude that using the weaker search criterion (b) alone has too many false positives to be practical for our purposes. Criterion (c) is only useful to look at in combination with criterion (a) or (b).

Database analysis

We extended the siRNAdb database by incorporating 776 HuSiDa data points classified as effective, resulting in a total of 1276 unique siRNAs. For consistency, we only used the 1188 19-mers. All potential database matches were generated for the human, mouse and rat siRNAs by performing an un-gapped WU-BLAST database search against the positive strand of the corresponding RefSeq database (Pruitt et al., 2005). Searches were performed with a word size of 1 and an E-value cutoff of 1000. Hits were parsed by MSPcrunch (Sonnhammer and Durbin, 1994) and subsequently elongated in either direction as required to generate a 19-bp ungapped alignment between the siRNA query and the target gene. The potential of each 19-mer match to elicit off-target effects was then examined using scoring methods described above to determine the likelihood of a non-specific effect being detected. For those cases where the accession number in the database is not a RefSeq accession number, we identified the corresponding accession number using a standard database search.

Hits to self were ignored, and we chose to filter out matches to closely related sequences (i.e. splice variants) where the hit region was enclosed in a high-scoring segment pair (HSP) of at least 80% identity, and the complete alignment covered at least 60% of the gene. These cutoffs ensure that all hits to splice variants are removed, at the potential expense of also removing some hits to other genes. Using these criteria reduced the number of siRNAs with matches of score 100 (perfect match) to 5, and the number of siRNAs with mismatches of score > 95 (one mismatch in a non-critical position) to 11 (table S1). In total 97 matches of score > 95 were found for 11 siRNAs. In three of these 11 cases, a total of 88 hits were found with a score > 95. This indicates that some siRNAs are highly non-specific.

The distribution of SS scores for experimentally verified siRNAs are shown in figure S2. We identified the best hit for each siRNA, ranked according to SS_{AVG} score. We selected a cutoff score of 80 for determining if an siRNA is predicted to have an off-target effect. A score of 80 or above represents ≤ 3 mismatches, predominantly in the terminal 3' and 5' positions (1,2,18,19). For example, a hit with mismatches at positions 18 and 19 gets a score of 92 (100-3-5). In contrast, a hit with mismatches at positions 5 and 10 would get a score of -34 (100-51-83), and is unlikely to elicit an off-target effect. SS_{AVG} scores of > 80 were observed for 16% of all cases, indicating a significant likelihood of observing an off-target effect.

The rules derived from microarray specificity data were also applied to the siRNAs. Table S2 shows the results from using these rules. For 32% (381/1188) of the siRNAs, at least one hit satisfying the most conservative (a) criterion was found, 90% (1069/1188) satisfied the less stringent (b) criterion, and 4% (51/1188) satisfied both criteria.

Detailed tables with all hits found are available as supplementary data at http://sirna.sbc.su.se/supplementary.html.

Discussion

We developed a novel scoring scheme based on experimental data to estimate the likelihood of an siRNA eliciting an off-target effect. Using this we were able to estimate the number of published siRNAs with significant off-target matches. By utilizing a number of scoring schemes we found that a high proportion (16-32%) of siRNAs reported in the literature have a high potential to cause off-target effects depending on the criteria selected. This figure would be much higher if we had included genes with high sequence identity to the query gene.

Close examination of hits with 100% identity revealed that almost all examples hit either different transcript variants of the same gene or closely related genes, hence these were excluded. These siRNAs may have been designed with potential off-target effects in mind. However, high-scoring but lower identity matches have the potential to affect genes that are not related, causing unexpected experimental results.

The figures presented here suggest less widespread non-specific effects than estimated by Snove and Holen (Snove and Holen, 2004), who found that 75% of 359 published siRNA sequences are likely to elicit an off-target effect. This may be explained by our treatment of mismatch positions as unequal, in addition to the use of a filter to ignore similar genes. If we included criteria (b) from the array-based scheme then we would obtain a figure around 90%, however this figure is a high upper bound.

Other scoring methods were considered but not used in this study. One such alternative strategy would be to use the stability (Tm) of the duplex formed. However, as with most thermodynamic approaches, it cannot take into account that the siRNA guide strand is part of a complex of proteins which may have an effect on the binding capabilities of the guide strand. Using this strategy would also fail to take into account the critical nature of certain nucleotide positions, as evident in figure S1 (such as the scissile nucleotides), although this could potentially be rectified by incorporating a weighting scheme based on the one we introduce here.

Supplementary References

Chalk, A. M., R. E. Warfinge, et al. (2005). "siRNAdb: a database of siRNA sequences." Nucleic Acids Res **33**:D131-4.

Du, Q., H. Thonberg, et al. (2005). "A systematic analysis of the silencing effects of an active siRNA at all single-nucleotide mismatched target sites." Nucleic Acids Res **33**:1671-7.

Jackson, A. L., S. R. Bartz, et al. (2003). "Expression profiling reveals off-target gene regulation by RNAi." Nat Biotechnol **21**:635-7.

Pruitt, K. D., T. Tatusova, et al. (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **33**:D501-4.

Snove, O., Jr. and T. Holen (2004). "Many commonly used siRNAs risk off-target activity." Biochem Biophys Res Commun **319**:256-63.

Sonnhammer, E. L. and R. Durbin (1994). "A workbench for large-scale sequence homology analysis." Comput Appl Biosci **10**:301-7.

Truss, M., M. Swat, et al. (2005). "HuSiDa--the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells." Nucleic Acids Res **33**:D108-11.