D Optimization Power Study

Due to the equation to calculate the T-Statistic, there is a close relationship between D and the eventual T value. Calculating the D for an entire model is relatively fast. What takes vastly more time is the calculation of the individual contributions to the D2. If we can minimize the number of times we delve into spending precious cycles calculating the D^2 for models that have no potential for being winning models, we can achieve times that are more in line with other analyses.

MDR-PDT has been configured to allow an optional optimization in which the program will abort the evaluation for D^2 if D fails to exceed a user defined percentage of the highest D observed so far. It is expected that setting the default to a reasonably high threshold will allow most analyses to be performed very quickly. If the results look questionable, they will be able to reduce the level to verify that the results hadn't been skewed by the optimization.

In order to determine the cost on power of using the optimization, a number of runs were performed at using the following levels of optimizations: 0% (no optimization), 50%, 75%, 80%, 85%, 90%, 95%, 98% and 99%).

Because the simulated datasets were very small (only 10 snps), the benefit is masked by the time it took to load the data and manage the results. In order to showcase the effect on real data, several datasets examined with the same variations and their ability to find the same model as 0% was observed, along with the difference in runtimes for each of the sets.

The data used was a subset of the data from the original MDR-PDT power study. The datasets included: 6 sets consisting only of trios whose children had 0% phenocopy error, 6 sets consisting only of triads whose affected children had 50% phenocopy error, and a single dataset with 2 affecteds and 1 unaffected, presumably at 50% phenocopy error.

Trio Data

Phenocopy does clearly interfere with MDR-PDT, which was reported in Martin, et al (2006). Optimizing the D does as well, however, it's affect on the 2 locus search resulted in an average loss of 1.58%. The average time to complete the search is difficult to see here due to roundoff from the timer. However, at 99%, it appears to be take about 25% of the time the unoptimized run took. According to this particular run, 80% represents a good choice for time vs. power.

Averaged Over All Sets

D Setting	Power	Average Time (s)	Overall Time (s)	Power Delta
0.00	61.73	0.04	96.51	0.00
0.25	61.73	0.04	94.48	0.00
0.50	61.92	0.03	83.63	0.18
0.75	61.75	0.02	69.42	0.02
0.80	61.72	0.01	67.07	-0.02
0.85	61.45	0.01	64.69	-0.28
0.90	61.08	0.01	62.6	-0.65
0.95	60.75	0.01	60.81	-0.98
0.98	60.15	0.01	59.88	-1.58
0.99	60.15	0.01	59.85	-1.58

Triads 3Snp Search

When performing 3 snp searches on trio data / no phenocopy, optimization at 80%, the runtime is approximately 20% of the unoptimized, and our ability to detect a model that included the preferred loci was actually increased marginally.

Averaged Over All Sets

D Setting	Power	Average Time (s)	Overall Time (s)	Pseudo Power	Pseudo Power Delta
0.00	4.62	0.3	375.69	99.33	0.00
0.25	4.48	0.29	362.81	99.33	0.00
0.50	4.22	0.15	221.57	99.33	0.00
0.75	5.35	0.19	148.83	99.35	0.10
0.80	5.03	0.06	138.41	99.35	0.10
0.85	4.35	0.16	128.97	99.3	-0.20
0.90	3.93	0.05	119.9	99.23	-0.60
0.95	4.9	0.04	111.56	99.1	-1.40
0.98	7.95	0.03	105.06	98.95	-2.30
0.99	9.62	0.03	104.35	98.95	-2.30

The power begins showing signs of degradation as the optimization is increased beyond 80%, with 99% showing only a 2.30% loss of power.

2 Affected / 1 Unaffected Data

The side effects of phenocopy on this set are higher than the average of the triad data due to the fact that we only had a single set- and the signal associated with it was second to the weakest of all of the simulated data. The effect from 0% to 99% is 2.1%. Again, 80% makes an excellent choice- ironically, it's power is actually higher than the unoptimized version despite it's runtime being significantly reduced. However, even at 99%, the delta is very small.

D Setting	Power	Average Time (s)	Overall Time (s)	Power Delta
0.00	44.90	0.05	95.35	0.00
0.25	44.90	0.04	92.72	0.00
0.50	45.20	0.03	78.39	-0.05
0.75	46.10	0.01	63.17	-0.20
0.80	45.50	0.01	60.85	-0.10
0.85	44.90	0.01	58.51	0.00
0.90	44.00	0.01	56.81	0.15
0.95	43.10	0.01	55.25	0.30
0.98	42.80	0.01	54.58	0.35
0.99	42.80	0.01	54.39	0.35

Real Data

5 Datasets were used to evaluate overall performance of the optimization. The same optimization settings were used. To compare success/failure, comparisons were made to the unoptimized run. In one comparison, the number of variations in the top five list were observed for each setting above 0%. In the other comparison, it was noted whether or not the top model from the unoptimized was lost.

Top Five Model Variations 3 SNP Search

	Maximum T	D Optimization Setting								
		0.25	0.5	0.75	0.8	0.85	0.9	0.95	0.98	0.99
Alzheimer's	8.6	0	0	0	0	0	0	3	5	5
Schizophrenia	5.68	0	0	2	4	5	5	5	5	5
BUX-350	5.09	0	0	0	0	0	0	1	1	2
Lupus	4.8	0	0	0	0	0	1	3	4	4
BPACDIT	3.89	0	0	1	3	3	4	5	5	5

* Red Cells indicate the configuration where the top model differed from the unoptimized run

Schizophrenia	4.59	0	0	0	1	2	2	4	4	4
BUX-350	3.63	0	0	0	0	0	0	1	1	1
Lupus	4.13	0	0	1	3	3	3	3	3	3
BPACDIT	3.56	0	0	1	1	1	1	2	2	2

* Red Cells indicate the configuration where the top model differed from the unoptimized run

Effects on Runtime

		Runtime Percentage								
	Initial Time (s)	**0.25**	**0.5**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**0.98**	**0.99**
Alzheimer's	280.29	0.84	0.1	0.08	0.06	0.03	0.02	0.01	0.01	0.01
Schizophrenia	207.07	0.99	0.5	0.07	0.05	0.04	0.03	0.03	0.03	0.03
BUX-350	14.31	0.97	0.61	0.13	0.09	0.06	0.04	0.03	0.03	0.03
Lupus	0.59	0.98	0.92	0.32	0.24	0.22	0.17	0.1	0.08	0.07
BPACDIT	10.6	0.84	0.25	0.14	0.14	0.12	0.11	0.11	0.11	0.11

Effects range widely from set to set. A lot of the variation is expected to be due to the arrangement of the SNPS. If one of the highest D values falls early on in the search, it will knock out a very large number of subsequent calculations. If those numbers steadily increase, the effects of the optimization will be much slighter.

However, even at 50% optimization, the amount of time saved is significant. With one exception, it is at 50% or better.