

Analyzing Blast-Like-Alignment-Tool (BLAT) for Gene Mapping.

By Olga Prilepova

Advisor: Rasmus Nielsen

One of the DNA research most promising avenues is the ability to decode one's personal DNA in a reasonable amount of time. This requires the ability to parse through the DNA of an individual and map it back to a DNA model even with significant differences between the two. This research was focusing on mapping the sequences produced by restriction enzymes cuts with BLAT (Blast Like Alignment Tool created by Jim Kent from UCSC) and analyzing the results.

For the purpose of this research we chose to mimic the results of the restriction enzyme ECOR1 on Human Genome. ECOR1 searches DNA molecule for "gaattc" sequence of nucleotides, then cuts the motif after "g" and before "aattc". This enzyme is commonly used in DNA sequencing. For our computer simulation Bash and C parsing scripts were used on the hg18 Human Genome database from UCSC website and the results were used as input to BLAT.

Running Bash/C scripts for "cutting" the DNA produced a large set of 778111 sequences. This set was far too big for analysis and contained many long sequences that would be simple to map. To ensure that we address the worst case scenario, we chose to use only the pieces no longer than 800 base pairs. These would provide the greater challenge for mapping, and thus produce a more reliable result in terms of the least number of sequences that can be mapped. The remaining set of 165305 was still too large for analysis, so the results were further narrowed by randomly selecting 10% (16402) of them for input into BLAT.

The BLAT program then mapped each sequence back to the original DNA molecule with a parameter of 97% sequence identity matching, in order to get the best hits, since human DNA variation should be between 1 and 2%, and the 3% mismatch is included for extra reliability. The results of this map would fall into 3 categories: those that were not mapped anywhere in the Genome, those mapped only once (uniquely) and those that were mapped to multiple places. Only a uniquely mapped sequence can be considered a success, since it gives the exact location of the sequence inside the genome. It was also important to see how well BLAT can perform on 1, 2 or 3% mismatch allowed.

Mutation	0%							
Sequence size	<=800bps							
Total number of cuts	165305							
Randomly chosen 10%	16402							
Allowed mismatch	0%		1%		2%		3%	
amount/percent	#	%	#	%	#	%	#	%
Not mapped	772	0.05	768	0.05	767	0.05	767	0.05
Mapped once	15338	0.94	11944	0.73	11772	0.72	11707	0.71
Multiple occurrences	292	0.02	3690	0.22	3863	0.24	3928	0.24

As it can be seen from the results of mapping sequences from the original 0% mutated DNA with BLAT and allowing for 0% mismatch the BLAT performance is quite high, since 94% of the sequences are uniquely mapped. The 5% of unmapped sequences is generally due to their short length, since the program is not meant for sequences of length less than 40 base pairs. Another 2% are repeats that may have come from different parts of DNA. It is reasonable to believe that increasing the allowed mismatch for BLAT results would increase the number of repeats and thus decrease the number of uniquely mapped sequences while the number of unmapped sequences should stay approximately the same, and that is exactly what the experiment is showing for this case, with unmapped cases staying around 5%, unique maps going down to 71% and repeats growing to 24%.

Also, since the individual DNA may be different that the hg18 model used as the database for BLAT search, 3 copies of the DNA were produced, with a random mutation of 1, 2 and 3% respectively, and the process of simulated enzyme cuts and BLAT map was performed again.

Mutation	1%							
Sequence size	<=800bps							
Total number of cuts	164742							
Randomly chosen 10%	16370							
Allowed mismatch	0%		1%		2%		3%	
amount/percent	#	%	#	%	#	%	#	%
Not mapped	9990	0.61	3858	0.24	1262	0.08	919	0.06
Mapped once	6204	0.38	9052	0.55	11394	0.70	11660	0.71
Multiple occurrences	176	0.01	3460	0.21	3714	0.23	3791	0.23

Mutation	2%							
Sequence size	<=800bps							
Total number of cuts	163634							
Randomly chosen 10%	16168							
Allowed mismatch	0%		1%		2%		3%	
amount/percent	#	%	#	%	#	%	#	%
Not mapped	11025	0.68	7458	0.46	3696	0.23	1645	0.10
Mapped once	5020	0.31	5598	0.35	8996	0.56	10906	0.67
Multiple occurrences	123	0.01	3112	0.19	3476	0.21	3617	0.22

Mutation	3%							
Sequence size	<=800bps							
Total number of cuts	162529							
Randomly chosen 10%	16052							
Allowed mismatch	0%		1%		2%		3%	
amount/percent	#	%	#	%	#	%	#	%
Not mapped	11392	0.71	8567	0.53	6492	0.40	3605	0.22
Mapped once	4549	0.28	4375	0.27	6110	0.38	8792	0.55
Multiple occurrences	111	0.01	3110	0.19	3450	0.21	3655	0.23

For the cases where mapped sequences are taken from mutated DNA the results can be interpreted as following: when the program allows for less mismatch than the variation of the input DNA, only some sequences can be mapped uniquely and many can't be mapped at all; when the program allows for the same percent of mismatch as the percent of mutation in input DNA, around 55% of sequences can be mapped and for more mismatch allowed – more sequences are mapped uniquely growing to approximately 70% , unmapped sequence number going down to ~10% and repeats growing to ~20%. The situation resembles the one of the original 0% mutated DNA.

It is interesting to note that only 55% of the sequences are mapped uniquely when the percent of allowed mismatch for each sequence is the same as the percent of mutation of the DNA from which the sequence came from, and this number grows as the allowed mismatch is increased. It would be quite logical to suppose that this happens due to the fact that when a sequence is of length less than 100 nucleotides and has mutation in it; the percent of mutation for that sequence is generally higher than the percent of mutation in the DNA it came from. (One mismatch in a sequence of length 99 is greater than 1%). If that is the case, then for effective (unique) mapping of sequences from input with variation, it is important to allow about 1 to 2% more mismatch than the percent of variation. For the individual Human Genome mapping it would be advisable to use 3% allowed mismatch.

Even though greater allowed mismatch produces better results, it is obvious that 100% mismatch would produce nothing useful, so it is important to find the upper bound for allowed mismatch that produces maximum uniquely mapped sequences. Thus it would probably be beneficial for the support of this hypothesis that an additional analysis of 4-7% (or more) allowed mismatch on BLAT results is done for the same data set.