Human-mouse alignments with Blastz

Scott Schwartz* W. James Kent† Arian Smit‡ Zheng Zhang§
Robert Baertsch† Ross Hardison ¶ David Haussler† Webb Miller*‖

September 12, 2002

Abstract

The Mouse Genome Analysis Group aligned the human and mouse genome sequences for a variety of purposes, using alignment programs that suited the various needs. For investigating issues regarding genome evolution, a particularly sensitive method was needed, to permit alignment of a large proportion of the neutrally evolving regions. We selected a program called Blastz, an independent implementation of the Gapped Blast algorithm specifically designed for aligning two long genomic sequences. Blastz was subsequently modified, both to attain efficiency adequate for aligning entire mammalian genomes and to increase its sensitivity. This paper describes Blastz, its modifications, the hardware environment on which we run it, and several empirical studies to validate its results.

1 Introduction

One of the goals set by the Mouse Genome Analysis Consortium (Mouse 2002) was to study mutation and selection, the main forces shaping the mouse and human genomes. Specific aims included (1) estimating the fraction of the human genome that is under selection (Roskin et al. 2002), (2) determining the degree to which genome comparisons can pinpoint the regions under selection (Elnitski et al. 2002) and (3) measuring regional variation in the rate and pattern of neutral evolution (Hardison et al. 2002). Attaining these goals required an alignment program with higher sensitivity than needed for other Consortium goals, such as predicting novel protein-coding segments or identifying large genomic intervals where gene order is conserved.

*Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802.
†University of California at Santa Cruz.
‡Seattle
§Paracel
¶Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802
‖Corresponding author. Tel: 814 865-4551; Fax:814 865-3176; Email: webb@bio.cse.psu.edu

Ideally, our alignment program would identify orthologous regions of the human and mouse genomes, whether or not they are under selection. That is, it would determine correspondences between genomic positions that are descended from the same position in the ancestral genome, allowing nucleotide substitutions. In practice, success in reaching that goal is measured by the program's sensitivity (fraction of orthologous positions that it aligns) and specificity (fraction of the aligned positions that are indeed orthologous). Most of our aims could have been addressed by a program that aligned neutrally evolving regions with a modest degree of sensitivity. For instance, regional variations (aim 3) could be assessed from a relatively small sample, (say, 1 out of 10 orthologous regions), provided that there were no critical biases in the sampling process. Demands on specificity were higher, but it was acceptable for, say, 5% of the aligned positions to be non-orthologous.

To meet out needs, we enhanced the Blastz alignment program (Schwartz et al. 2000). Here we describe the alignment program, the hardware environment, and several validation studies. Our results indicate that we have correctly determined the majority of what can be aligned between the human and mouse genomes.

2 Results

2.1 Software Design Issues

Our goal was to align an appreciable fraction of the neutrally evolving DNA in the human and mouse genomes. This sensitivity requirement disqualified several existing programs (Ning et al. 2001, Kent 2002) that sacrifice sensitivity to attain very short running times. Preliminary studies indicated that an appropriate level of sensitivity and specificity was attained by a program called Blastz, which is used by the PipMaker webserver (Schwartz et al. 2000) to align genomic sequences. Blastz follows the three-step strategy used by Gapped Blast (Altschul et al. 1997): find short near-exact matches, extend each short match without allowing gaps, and extend each ungapped match that exceeds a certain threshold by a dynamic programming procedure that permits gaps.

Two differences between Blastz and Gapped Blast were exploited in our whole-genome alignments. First, Blastz has an option to require that the matching regions that it reports must occur in the same order and orientation in both sequences. Second, Blastz uses an alignment scoring scheme derived and evaluated by Chiaromonte et al. 2002. Nucleotide substitutions are scored by the matrix

```
     A    C    G    T
A   91 -114  -31 -123
C -114  100 -125  -31
G  -31 -125  100 -114
T -123  -31 -114   91
```

and a gap of length k is penalized by subtracting 400 + 30k from the score. To determine whether an ungapped alignment is sufficiently promising to warrant initiation of a dynamic

programming extension step, the sum of scores for its columns is multiplied by a value between 0 and 1 that measures sequence complexity, as described in detail by Chiaromonte et al. This makes it harder for a region of extremely biased nucleotide content to trigger a gapped alignment.

Two changes to Blastz significantly improved its execution speed for aligning entire genomes. First, when the program realizes that many regions of the mouse genome align to the same human segment, that segment is dynamically masked, i.e., marked so that it will be ignored in later steps of the alignment process. Second, we adapted a very clever idea of Ma et al. 2002 for determining the initial short match that may seed an alignment. Formerly, Blastz looked for identical runs of 8 consecutive nucleotides in each sequence. Wu et al. propose looking for runs of 19 consecutive nucleotides in each sequence, within which the 12 positions indicated by a "1" in the string

```
1110100110010101111
```

are identical. To increase sensitivity, we allow a transition (A-G, G-A, C-T or T-C) in any one of the 12 positions.

Later, Blastz was further modified to utilize the increasing contiguity of the available mouse sequence. For the earliest alignments, the mouse sequence was presented in unassembled "reads" of around 500 bp each. An ungapped alignment was required to score at least 3000 (relative to the above substitution scores as adjusted by the measure of sequence complexity) before the dynamic programming step was initiated. Such a high threshold was needed to attain reasonable specificity. Availability of longer mouse segments suggested looking for pairs of alignments that connect nearby regions in both human and mouse genomes; the alignment procedure can be run with a lower threshold in the regions between the alignments. If the separation between the two alignments is under 50 kb in both sequences, then Blastz recursively searches the intervening regions for 7-mer exact matches and requires a threshold of 2200 for initiating dynamic programming (with the adjustment for sequence complexity). If the separation is under 10 kb, the threshold is lowered to 2000. In either case, the higher-sensitivity matches are required to occur with an order and orientation consistent with the stronger flanking matches.

Although the fee for initiating a gapped alignment is steep (e.g., 3000), once started the alignment keeps extending as long as the average score of the added alignment remains positive. This observation suggests a strategy of physically removing lineage specific interspersed repeats (i.e., that inserted after the human-mouse split), earlier utilized by Lee et al. (1998). Then, an alignment that is initiated on one side of the insertion point can freely jump the other side, where it may detect alignable nucleotides that may not meet the steep initiation fee on their own. We now remove recent repeat elements, run Blastz, then adjust positions in the alignment to refer to the original sequences. These last two improvements, i.e., recursive application of Blastz between adjacent alignments and excision of lineage specific repeats, increased alignment coverage from 32% of the human genome to 40%.

The modified Blastz was used to compare all the human sequence with all of the mouse, i.e., to produce a complete catalog of matching regions. Frequently, more than one region

of the mouse sequence aligned to the same region of the human sequence. This is a natural consequence of duplications in the mouse genome and in the human/mouse common ancestor. These duplications include paralogous genes, processed and unprocessed pseudogenes, tandem repeats, simple repeats, etc. For many purposes one wants the single best, orthologous, match for each human region. Typically when looking at a region spanning several thousand bases it is clear which alignment is the ortholog and which are the paralogs. The orthologous alignment usually is longer, and overall has a greater sequence identity. On the other hand, over a small regions by chance a paralog may have greater sequence identity than an ortholog.

We wrote a program, axtBest, which automatically picks out an alignment likely to be the orthologous alignment. The program works in two passes. A window of 5000 bases on either side of a human base is used to score that base. The score is the Blastz score at the end of the window minus the Blastz score at the beginning of the window. Bases less than 5000 bases from the start or end of the alignment have to be handled specially, and the details of this differ between the first and the second pass. In the first pass the window is truncated to only cover the alignment itself. This tends to make bases on the edge of an alignment or bases in a short alignment score less than bases in the middle of a long alignment. Unless an alignment is the best scoring alignment over at least one human base in the first pass, it is thrown out. In the second pass bases on the edges of an alignment are given the same score as the base 5000 bases into the alignment. The portions of each alignment remaining after the first pass that are the best scoring over at least 10 bases of the human genome are then written out. This two pass scheme with the first pass being done on softer-edged windows and the second pass on harder-edged windows in most cases yields the same result one would obtain with a simpler one pass scheme. However when applied to regions with multiple paralogs it results in less fragmented 'best' alignments than the single pass schemes we have tried.

2.2 Implementation Issues and Hardware Environment

We precompute a list of jobs, each aligning 1.01 Mb of human sequence with 30 Mb of mouse, allowing 10 kb overlaps between adjacent human segments. The scheduling software available on the 1024 CPU cluster that we use doesn't support processor affinity of any sort; each job is entirely independent. Oblivious scheduling makes it unattractive to reduce disk space requirements by devoting each node to a particular genomic region, which would involve significant administrative overhead. Fortunately, we have sufficient local disk on each node to provide a copy of all the input that might be required. Only the output needs to be communicated to a shared resource, and this is the main I/O bottleneck in our procedure, due to the latency of using a network file system.

Dynamic masking was invented to handle cases like human chromosome 19, where zinc finger or olfactory receptor genes match a huge number of times, but are not flagged as repeats. Oblivious scheduling partially defeats this optimization, however, since each human segment is compared many times to different parts of mouse, and no efficient mechanism is available to share dynamic masking information.

To align 2.8 Gb of human sequence vs. 2.5 Gb of mouse sequence took 481 hours of CPU time and a half day of wall clock time on a cluster of 1024 833 Mhz Pentium III CPUs. This produced 9 Gbytes of output in a relatively space-efficient format that describes the alignments by coordinates within the sequences. These are translated to a textual representation, called axt, that includes the actual base pairs. While axt files are large, for many post-processing steps the improved locality of reference (avoiding the need to retrieve parts of multi-gigabyte datasets) is a clear necessity. The initial axt files were 20 Gbytes, but running axtBest reduced them to 2.5 Gbytes. Only 3.3% of the human genome is covered by multiple alignments, but some of these places, particularly on chromosome 19, are covered to a great depth.

2.3 Software Evaluation

Like other alignment programs, Blastz permits adjustment of numerous parameters and thresholds, such as gap penalties and the threshold of 3000 for initiating a gapped alignment based on a gap-free alignment. It is a tricky business to test whether a particular combination of values is doing a good job. Many other classes of sequence analysis software benefit from availability of an experiment-based gold standard; protein alignments are checked against X-ray crystal structures and gene predictions are compared to cDNA sequences. On the other hand, when trying to correctly align neutrally evolving sequences, it is not clear how one can determine the correct answer. The maxim "it is an order of magnitude easier to design two good programs than to tell which one is better" seems appropriate here.

Our aim was to determine all homologous matches between human and mouse genomic regions, which are then filtered and examined in a variety of ways for a variety of purposes, such as identifying regions of conserved synteny. Early in the project, when we were working with unassembled mouse reads, there was no choice but to compare all of the human sequence with all of mouse. Later, when reliable assemblies were produced and syntenic segments and blocks had been identified at moderate resolution, it became feasible to avoid an all-vs-all computation (though this would preclude some of the studies that we want to perform). Computational efficiency would increase dramatically if each region of the human sequence were compared to a small segment of the mouse genome rather than to all of it. Another benefit would be to substantially decrease the likelihood of a match to unrelated sequence occurring by chance.

However, experimental evidence indicates that the level of spurious matches in our all-vs-all comparisons is quite low. We reversed the soft-masked mouse sequence (without complementing it) and aligned it to human. The reversed mouse sequence has precisely the same size and local compositional complexity as mouse, for instance, a micro-satellite sequence cacaca... in mouse is preserved in the reversed sequence. Thus, the quantity of matches to the reversed mouse should accurately reflect the quantity of spurious matches to mouse. While about 40% of the human sequence aligned to mouse, only 0.16% aligned to reversed mouse, confirming high specificity of our approach even before axtBest is applied.

Extra benefits accrued from this experiment. For instance, it indicated that discarding

regions of the human genome aligning to more than one region in mouse would eliminate nearly half of improperly aligned human regions while retaining 94% of the correctly aligned ones. Another observation is that the sensitivity/specificity tradeoff can be improved by minor adjustments to the outer alignment step. For instance, imposing the requirement that gapped alignments score at least 5000 reduced outer-alignment coverage by only 0.2% of human, but lowered coverage by bogus alignments from 0.16% to 0.07%. Also, our measurements distinguished "outer alignments" (all-vs-all) from "inner alignments" (higher sensitivity searches between two adjacent outer alignments). Blastz's inner alignment steps searched 26% of the human sequence and added 2% of it to reported alignments. The ratio of true positives to false positives was about 1000:1, suggesting that sensitivity in inner alignment steps can be safely increased. The average length of a human region searched during an inner-alignment computation was 3 kb, hence the "alignment space" for the all-vs-all outer alignments is 3000 times bigger than for inner alignments (2.9Gb-by-2.5Gb vs. 0.8Gb-by-3kb). This suggests that inner alignments can be performed at relatively low efficiency, say, with a Smith-Waterman algorithm, without appreciably affecting the total computation time.

Another test of the specificity of Blastz alignments is based on synteny. Human chromosome 20 is considered to be completely syntenic to parts of mouse chromosome 2. After filtering through axtBest only 3.3% of the human chromosome 20 bases in alignments aligned outside of mouse chromosome 2. A certain degree of non-syntenic alignment is to be expected from processed psuedogenes. Since only approximately 96% of the mouse genome is sequenced, in some cases a paralog on another chromosome will fill in for a non-sequenced ortholog on chromosome 2 as well.

3 Discussion

The Mouse Genome Analysis Group employed a variety of programs to align the mouse and human sequences, including Blast (Altschul et al. 1997), Blast2sequences (Tatusova and Madden 1999), BLAT (Kent 2002) and PatternHunter (Ma et al. 2002). However, none of these programs was suited to our goal of investigating fine-scale features of genome evolution, primarily because they were tuned for aligning protein-coding regions, whereas our focus was neutral evolution. Hence, we chose to develop a new program. An empirical comparison of available tools indicated that the Blastz program was a good place to start.

Two aspects of Blastz's design philosophy proved valuable. First, Blastz is intended for use at all stages of genome sequencing, including the initial period when only small contigs, or even simply unassembled 500-bp reads, are available. Blastz permitted us to begin our analyses quite early. This capability is of course essential in cases where sequencing of a second genome stops with light shotgun coverage.

Another tenant of the Blastz philosophy is that the alignment program should not enforce critical a priori assumptions about which alignments are important; rather, it should be fairly inclusive. The task of processing and filtering the initial alignments in various ways is left to down-stream programs, which can be made quite flexible and efficient (e.g., Huang

et al. 1994, Zhang et al. 1999).

We expect that further work on alignment programs will soon yield a 10-fold reduction in execution time for all-vs-all genome comparisons with no degradation of sensitivity and specificity over Blastz. For finished mammalian genomes, reliable coarse-grained mapping of homologous regions will make it possible to update a genomic alignment on a single workstation. All this can be accomplished using the Gapped Blast design. Different approaches may do even better.

A number of difficult tasks remain before the problem of aligning two mammalian genome sequences is adequately solved. Echoing the sentiments of Miller (2001), we look forward to progress in the intertwined areas of producing higher-sensitivity alignments and of evaluating their correctness, and hope that investigators will adopt a cooperative spirit reflecting the highest ideals of the Genome Project.

4 Acknowledgments

We greatly appreciate the feedback given by Krishna Roskin, Ryan Weber, Angie Hinrichs and Mark Diekhans. S.S., R.H. and W.M. are supported by grant HG-02238 from the National Human Genome Research Institute.

5 References

Altschul, S. F., Madden, T.L., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST – a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Chiaromonte, F., Yap, V.-B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. *Pacific Symposium on Biocomputing,* 115–126.

Elnitski, L., Hardison, R. C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M. J., Schwartz, S., Miller, W., and Chiaromonte, F. 2002. Distinguishing regulatory DNA from neutral sites. Mouse satellite paper submitted to *Genome Research.*

Hardison et al. 2002. Co-variation in divergence by substitution, deletion, transposition and recombination during mammalian evolution. Mouse satellite paper submitted to *Genome Research.*

Huang, X., Pevzner, P., and Miller, W. 1994. Parametric recomputing in alignment graphs. *Combinatorial Pattern Matching* (Springer Lecture Notes in Computer Science, 807), 87–101.

Kent, W. J. 2002. BLAT – the BLAST-Like Alignment Tool. *Genome Research* **12**, 656–664.

Lee, I.Y., Westaway, D, Smit, A.F., Wang, K., Seto, J., Chen, L., Acharya, C., Ankener, M., Baskin, D., Cooper, C., Yao, H., Prusiner, S.B., and Hood, L.E. 1998. Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome Research* **8**, 1022–1037.

Ma, B., Tromp, J., and Li, M. 2002. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445.

Miller, W. 2001. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391-397.

Mouse (2002) Initial sequencing and comparative analysis of the mouse genome. Submitted to *Nature*.

Ning, Z., Cox, A J., and Mullikin, J. C. 2001. SSAHA: a fast search method for large DNA databases. *Genome Research* **11**, 1725–1729.

Roskin, K., et al. 2002. Satellite paper on fraction under selection. Submitted to *Genome Research*.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker – a Web server for aligning two genomic DNA sequences. *Genome Research* **10**, 577–586.

Tatusova, T.A., and Madden, T.L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* **174**, 247–250.

Zhang, Z., Berman, B., Wiehe, T., and Miller, W. 1999. Post-processing long pairwise alignments. *Bioinformatics* **15**, 1012–1019.