Charlie Liu MCDB 452 Term Paper 12/12/06

Human-Bovine Synteny Mapping

In the past decade, bovine genome research focused on linkage mapping and finding economic trait loci (ETL). Systematic application of mapped DNA markers to pedigree analysis has successfully discovered traits of interest. In order for these searches to be complete and the results efficiently applicable to selective breeding, the specific genes responsible for economic traits must be identified and cloned. Given the relatively sparse funding devoted to bovine genome research, the most practical approach to finding candidate genes would be comparative mapping with the much more extensively studied human and mouse genomes. By 1995, the boundaries of conserved human-bovine genome synteny had been fairly well defined through comparative mapping on human chromosomes and chromosome painting on bovine chromosomes. Within these boundaries; however, conservation of gene order remains undefined. A review paper that year pointed out the necessity of ordering evolutionarily informative markers, and the authors proposed the use of comparative mapping anchor loci for the task (Womack and Kata 1995). This line of research seems to have been abandoned—a quick search with Google Scholar reveals no papers on the topic since 1998—despite the fact that recent developments in synteny alignment make the task much more feasible. I propose the application of recent comparative mapping anchor techniques to examine conservation of gene order within conserved syntenies in humans and bovines.

Currently, the most popular pairwise sequence alignment algorithms employ either dynamic programming or hash tables. Dynamic programming was first applied to sequence alignment by Needleman and Wunsch in 1970. Their method for global alignment iteratively optimized amino acid sequence similarity based on a user-defined scoring matrix, with specific scores assigned for each pairwise amino acid match and gap opening and extension penalties. This quantified scoring system allowed not only alignment optimization, but also meaningful distance comparisons between different sequence pairs (Needleman and Wunsch 1970). A subsequent paper proposed a similar algorithm for segment pairs, which would allow for insertions and deletions of any length. While this algorithm is better suited for local alignments, it is more computationally intensive (Smith and Waterman 1981). Even with the use of dynamic programming, the computation involved in optimizing the comparison matrix for both methods makes aligning long sequences and querying a sequence against a database prohibitively slow.

To speed the process, sequences are compared using hashing algorithms, where sequences are not compared on an individual amino acid basis, but divided into ordered "ktuples" of user defined length and used to construct lookup tables. The ktuple is treated as a block, effectively dividing the sequence length by k as far as processing time is concerned. All elements of a query and database ktuple must align for that ktuple to count as a match. Consequently, larger ktuples make for faster, but less sensitive, searches. The FASTA algorithm uses hashing and iterative joining of the ten highest scoring diagonals to produce fast, selective alignments (Pearson 1990). The BLAST algorithm extends hash hits into high scoring segment pairs. Extension stops when the score stops improving, the best scoring segment against each database sequence is found, and their score significance is determined statistically (Altschul et al. 1990). A subsequent gapped BLAST allows for gapped alignments, and also employs a "two hit" method which increases hash hits, but reduces the number of segment extensions performed. Because extension is the slowest step of the BLAST algorithm, gapped BLAST is three times faster than the original. A new algorithm, PSI-BLAST, iterates alignments based on position-specific score matrices derived from the previous round of gapped BLAST alignment. The iterative nature of PSI-BLAST makes it slower but more sensitive, and thus useful for detecting protein families (Altschul et al. 1997).

The standard local sequence alignment tools above compare two genomes without assuming any specific evolutionary relationship between them a priori. Syntenic alignment, on the other hand, aims to map orthologous blocks—regions assumed to derive from the same genomic region of the two species' nearest common ancestor. Unfortunately, local alignment alone is an insufficient predictor of synteny because local alignments between non-orthologous locations are abundant. These false positives mainly arise from sequence duplications and intragenomic repeating regions. Gene loss and rearrangement and sequence divergence and deletions further complicate synteny detection. The complications also underscore the synteny map's importance as a tool for sorting out true orthology from the enormous amount of false positive noise that simple local sequence alignment yields (Batzoglou 2005).

The authors of the mouse genome sequence paper in 2002 performed a large scale synteny mapping using their robust draft of the mouse genome sequence and the >90% finished human sequence. The authors defined orthologous landmarks as regions where the sequences exclusively match over at least 40-bases perfectly. These landmarks are referred to in subsequent papers and elsewhere in this paper as "anchors." The authors found 558,000 anchors. These anchors comprised about 7.5% of the mouse genome with a mean spacing of about 4,4 kb in the mouse assembly. These anchors were compared between human and mouse genomes to identify syntenic segments, which the authors defined as maximal regions in which anchors occur in the same order on a single chromosome in both species. To minimize false positive noise resulting from imperfections in the draft genomes, the authors limited their search to regions over 300 kb. Even with this condition, over 90% of both the human and mouse genomes unambiguously resided within conserved syntenic segments. The resulting synteny map showed extensive intra- and inter-chromosomal rearrangement in most human and mouse chromosomes (Waterson et al. 2002).

Using the same set of anchors as Waterson et al., Pevzner and Tesler implemented an algorithm known as Genome Rearrangements in Mammalian Evolution (GRIMM) to construct human-mouse "synteny blocks" from the sequence alignment. The authors defined synteny blocks as segments that could be rearranged into conserved segments. They are of interest because they are the potential result of micro-rearrangements of conserved segments. To find these blocks, the authors concatenated the human and mouse genomes to form a Cartesian coordinate system. The anchors were plotted, and all anchors with Manhattan distances below a threshold were connected by an edge. The resulting connected components, known as "clusters," were discarded if their length failed to exceed a second threshold. Synteny blocks were formed from the strips in the remaining clusters. While syntenic blocks allow for the examination of evolutionary rearrangement, only the longer ones are reliable because spurious local alignments and sequencing errors may be responsible for shorter blocks (Pevzner and Tesler 2003).

A third method used the local alignment tool BLAT, an offshoot of BLAST (Kent 2002) to find anchors and AVID (Bray et al. 2003) for global alignments. The anchors were then grouped by proximity, order, and orientation, and the strongest group of anchors was used to map each mouse contig to a human complement. Of the three, this method employs the most updated and powerful tools available (Couronne et al. 2003). All three methods closely agreed on long syntenic segments. They primarily differed in syntenic resolution and ability to handle local shuffles (Batzoglou 2005).

In 2004, Gibbs et al. published a draft sequence of the rat genome, providing opportunity for synteny detection across multiple mammalian genomes (Gibbs et al. 2004). The GRIMM methodology (Pevzner and Tesler 2003) was modified to find all pairwise and three-way blocks amongst the human, mouse, and rat genomes (Bourque et al. 2004). Two additional mapping methods were developed to deal with all three genomes. Pash implements positional hashing, a novel method that divides hash tables along diagonals for simultaneous processing, allowing for reasonably large alignments of long sequences without sacrificing sensitivity. Unlike BLAST and BLAT (Altschul et al. 1997; Kent 2002), whose processing times are based on the product of sequence lengths, Pash processing time is a function of the sum of sequence lengths, making Pash a better choice for aligning long sequences and determining anchors for whole genomes (Kalafus et al. 2004).

Meanwhile, Brudno et al. developed a two step progressive alignment. They first aligned mouse and rat genomes; then aligned the human genome to either the mouse-rat alignments or to the remaining unaligned sequences. The authors then used the methods of (Couronne et al. 2003) to derive a synteny map, which they compared with one independently based on gene predictions to verify the alignment's ortholog-mapping accuracy (Brudno et al. 2004).

As modern synteny mappers have observed, micro-rearrangements abound between the human and mouse genomes (Waterson et al. 2002; Pevzner and Tesler 2003; Couronne et al. 2003), and obfuscate alignment within a syntenic block. This makes them an important issue for global alignment, and several methods have been developed to deal with the issue. These methods should be applicable to the bovine genome as well, and will hopefully shed light on gene order conservation within bovine-human syntenies.

Fundamentally, global alignments are less prone to false positives, while local alignments can better detect non-syntenic orthology. The concept of *glocal* alignment was introduced as a compromise between unfiltered local and strict global alignments. The authors present a glocal map as a chain of local alignments in which one sequence is monotonic, while the second sequence need not be. This allows the alignment to compensate for chromosomal mutation events like local inversions and transpositions. The algorithm for this method, Shuffle-LAGAN, combines previous global and local alignment algorithms. Shuffle-LAGAN computes the single monotonic sequence chains and then globally aligns its consistent parts. This procedure; however, does not properly model the boundaries of rearranged segments (Brudno et al. 2003).

The same year, an alterative compromise method between local and global alignments was proposed. The method, known as the chain-and-net approach, uses the program CHAINNET, which produces local alignments of monotonic chains from both sequences. Chains are then iteratively chosen and added to the global map. Parts of the chain that overlap with previously accepted chains from a genome of interest are thrown out, ensuring a one-to-at-most-one mapping if nucleotides from the genome of interest with those of the second genome. Shuffles; however, may occur in any size. The chain-and-net technique has a great advantage over GRIMM (Pevzner and Tesler 2003) in its ability to account for duplication and deletion in addition to transposition and insertion. GRIMM; however, can deal with translocations at chromosome ends, whereas the chain-and-net technique cannot handle overlapping rearrangements well and can only correctly classify translocations inserted in the middle of a chromosome (Kent et al. 2003).

This method revealed a very high frequency of rearrangements in the mouse genome relative to man. On average, every Mb of mouse genome had 2 inversions, and the median size for all such inversions was 814 kp (Kent et al. 2003). While the regional frequencies of these events are not uniform at all across the genomes (Batzoglou 2005), and the bovine genome is more similar to the human genome than the mouse is (Womack and Kata 1995), the high levels of inversion and other micro-rearrangement between both human and mouse and human and rat genomes (Brudno et al. 2004) strongly suggest that micro-rearrangements will occur on many syntenically aligned human-cow chromosome pairs as well.

Research last decade believed it had well-defined the boundaries of human-bovine synteny (Womack and Kata 1995). However, given the analogous situation of the mouse genome, in which recent techniques and new findings overturned prior beliefs (Couronne et al. 2003), even the “well-defined” aspects of human-bovine synteny may be worth re-examining with modern tools. A variety of algorithms, including strict thresholds, BLAT, and Pash (Waterson et al. 2002; Kent 2002; Kalafus et al. 2004), are now commonly used solutions to the exact anchor detection problem posed by Womack and Kata in 1995. Furthermore, the development of methods such as GRIMM, Shuffle-LAGAN, and CHAINNET (Pevzner and Tesler 2003; Brudno et al. 2003; Kent et al. 2003) allows for direct examination of gene order in human-bovine syntenies. Because the bovine genome is incomplete, the main purpose of synteny mapping in this case would be to make gene predictions based on extrapolations from local homology. Consequently, CHAINNET and GRIMM would be my algorithms of choice: CHAINNET for its ability to account for duplication and deletion, and GRIMM for its high level of sensitivity. While I would most heavily weight the CHAINNET results in my model, I would also use the Shuffle-LAGAN algorithm independently to compare and corroborate results based on the other methods. For my anchors, I would use either strict thresholds or BLAT, which would be more efficient than Pash for querying a series of shorter sequences (the very incomplete bovine genome) against a far larger database (the human sequence). The tools now exist to address the problem Womack and Kata posed in 2005, and the answer is still relevant.

References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* **215:**403–410, 1990.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25:** 3389-3402, 1997.

Batzoglou S. The many faces of sequence alignment. *Briefings in Bioinformatics* **6:** 6-22, 2005.

Bourque G, Pevzner PA, Tesler G. Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons from Human, Mouse, and Rat Genomes. *Genome Research* **14:** 507-516, 2004.

Bray N, Dubchak I, Pachter L. AVID: a global alignment program. *Genome Research* **13:**97–102, 2003.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S, LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**:721–31, 2003a.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. Glocal alignment: finding rearrangements during alignment. Special Issue on the Proceedings of the ISMB 2003, *Bioinformatics* **19:**54i-62i, 2003b.

Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglou S, Dubchak I. Automated Whole-Genome Multiple Alignment of Rat, Mouse, and Human. *Genome Research* **14:**685–692, 2004.

Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin, E, Pachter L, Dubchak I. Strategies and Tools for Whole-Genome Alignments. *Genome Research* **13:**73-80, 2003.

Dewey CN, Pachter L. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Human Molecular Genetics* **15:**R51-R56, 2006.

Gibbs RA, Weinstock GM, Metzker ML et al. Genome sequence of the Brown Norway rat yields insight into mammalian evolution. *Nature* **428:** 493-521, 2004.

Kalafus KJ, Jackson AR, Milosavljevic, A. Pash: Efficient genome-scale sequence anchoring by positional hashing. *Genome Research* **14:** 672-678, 2004.

Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Research* **12:** 656–664, 2002

Kent WJ, Baertsch R, Angie Hinrichs A, Miller W, Haussler D. Evolutions cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *PNAS* **20:**11484–11489, 2003.

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48:**443-453, 1970.

Pearson WR. Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Methods in Enzymology* **183:**63-98, 1990.

Pevzner PA, Tesler G. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomic sequences. *Genome Research* **13:**13–26, 2003.

Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology* **147:**195–197, 1981.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:**520–562, 2002.

Womack JE, Kata SR. Bovine genome mapping: evolutionary inference and the power of comparative genomics. *Current Opinion in Genetics & Development* **5:**725-733, 1995.