

Homology mapping with Markov random fields

Anat Caspi,¹ Lior Pachter,²

Keywords: Homology, ferromagnetic ising model, Markov networks, Markov random fields, multiple sequence alignment.

1 Background.

Evolution through divergence gives rise to different, though related, present-day genomes that shared common ancestors. Portions of genomes could be seen as genomic entities spawned through some dynamic changes in content and order of the ancestral genome. Certain regions, through selection, are conserved over time. Such genomic portions (be they gene-coding regions, conserved non-coding regions, etc.) that are related due to their derivation from the same element in a common ancestral genome are termed *homologs*. Comparative genomic techniques aim to infer genome organization and structure, as well as the evolutionary mechanisms that shaped present day genomes. Since much of our understanding of evolution comes from a comparative framework, it is important to accurately identify *homologs* in order to compare components that are actually linked by common ancestry.

The challenge in pairwise and multiple sequence comparison is to develop methods that are both biologically relevant (i.e., that account for the multitude of known evolutionary mechanisms) and that are also statistically sound (i.e., there should be a meaningful framework within which to compare alignments, and with which to evaluate the significance of an alignment). Furthermore, methods need to be practical for whole genome comparison.

2 The Homology Assignment Problem.

The recognition of homologous components requires two separate steps: (1) **sequence matching**: identifying similar sequence elements between genomes; (2) **homology mapping**: separating matches into homologous components and chance occurrence of similarity.

Alignment methods such as BLAST, BLAT and BlastZ focus on sequence matching, not homology mapping[1, 9, 11]. Typically, exact matches are identified using suffix trees, and are then extended locally to form alignments that are constrained to preserve the sequence ordering. The number of local alignments produced depend on the parameters of seed selection, but no explicit attempt is made to impose consistency among the matches or produce coherent homology maps. Figure 1 displays the well-known Cystic Fibrosis Transmembrane regulator (CFTR) example in four currently sequenced vertebrate genomes. Horizontal shaded bars represent the CFTR region in a fully sequenced chromosome. The lines between chromosome bars connecting small segments represent BLAT[9] output matches.

Though the methods use evolutionary point mutation models for generating locally good alignments, the original seeds are frequently not homologous (see spurious matches in figure 1). **Homology assignment** is the necessary selection mechanism to choose only those matches that are both *good sequence alignments*, and *preserve our preference for coherence within their local genomic context*. Specifically, by *coherency* we mean general preservation of order and orientation among nearby matches with rearrangement where necessary to maximize collinearity. Recent *anchor based* approaches ([8, 3, 5, 7]) obtain local alignments using BLAST-like methods, and provide a subsequent homology mapping step applied to the local alignments using dynamic programming algorithms or heuristics. The methods produce pairwise global alignments. Some extend to multiple alignment. **Shuffle-LAGAN** segments two sequences into a globally coherent alignment that locally preserves collinearity[6].

¹JGG in Bioengineering, University of California, Berkeley. caspian@compbio.berkeley.edu

²Department of Mathematics, University of California, Berkeley. lpachter@math.berkeley.edu

Figure 1: BLAT output

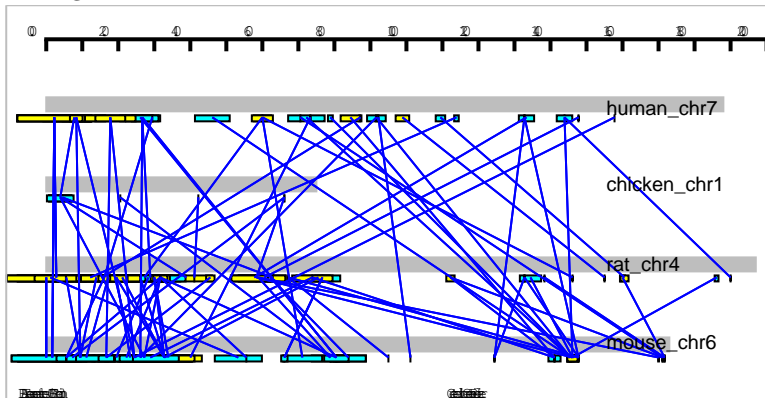
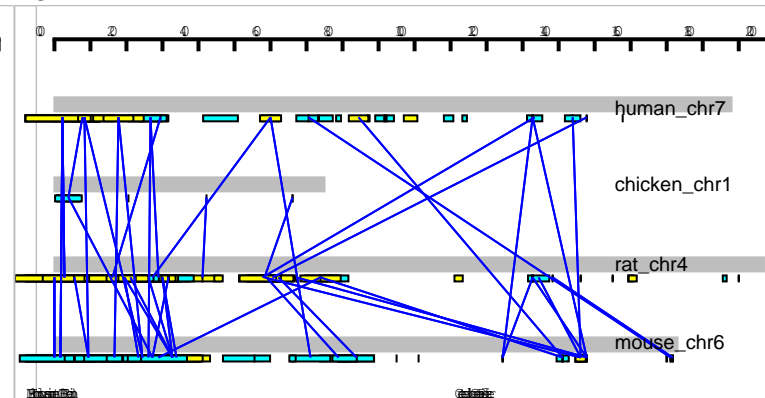


Figure 2: our MRF MAP output



Unfortunately, anchor-based methods *do not admit a probabilistic interpretation and rely on heuristics*. Furthermore, they mix the sequence matching and homology mapping steps, making it difficult to formulate rigorous results about their success in identifying homologous components. Methods such as **Shuffle-LAGAN** which allow for genome rearrangements, are difficult to extend to multiple alignment.

3 Methods and Results.

In this paper we propose a novel framework for multiple sequence comparison which expands the model of mutation, insertion and deletion to include operators for chromosomal duplication, inversion and micro-rearrangements. Insofar as sequence comparison is a search for homologous components in the sequences being compared, emphasis is placed on evaluating matches using their degree of similarity as well as their genomic context. The key to our approach is the probabilistic modeling of homology using Markov random fields. This allows us to define a probability distribution over the range of possible homolog mappings. We then formulate the homology assignment problem as a labeling problem for a special class of Markov networks known as the ferromagnetic Ising model[2]. This allows us to take advantage of recently developed efficient linear programming algorithms for exact inference[10].

We delineate the sequence matching from the homology mapping aspects of alignment, and outline the following approach to homology mapping: (1) Obtain matches (not necessarily identical) between multiple sequences in advance of the homology mapping: these matches may be pairs of single base pairs, larger BLAST hits, exons, or even complete genes. (2) Construct a constrained Markov random field based on the matches. (3) Find the MAP assignment using linear programming. (4) Output homologous matches.

We show that our approach compares favorably with existing mapping methods by analyzing a well-characterized region from Arabidopsis, Tomato and Capsella, as well as the CFTR region. Our method provides a tractable approach to multiple sequence homology assignment that is guaranteed to infer the optimal assignment (for given parameters) without resorting to iterative or progressive heuristics.

References

- [1] Altschul S et. al.: Basic local alignment search tool. J Mol Bio 1990, 215:403-410.
- [2] Besag J: The statistical analysis of dirty pictures. Royal Statistical Society 1986, B48/3:259-302.
- [3] Bray N, Pachter L: The MAVID multiple alignment server. NAR 2003, 13:3525-3526.
- [4] Bray N, Pachter L: MAVID. GR 2004, 14:693-699.
- [5] Brudno M et. al.: LAGAN and Multi-LAGAN. GR 2003, 13:721-731.
- [6] Brudno M et. al.: Glocal alignment. Bioinformatics 2003b, 19 Suppl 1:I54-I62.
- [7] Delcher A etl al.: Alignment of whole genomes. NAR 1999, 27:2369-2376.
- [8] Hohl M et. al.: Efficient multiple genome alignment. Bioinformatics 2002, 18 Suppl 1:S312-S320.
- [9] Kent J: BLAT- The BLAST like Alignment Tool. Genome Biology 2002, 12(4):656664.
- [10] Kolmogorov K, Zabih R: Multi-camera Scene Reconstruction via Graph Cuts. ECCV 2003.
- [11] Schwartz S et. al. : Human-mouse alignments with BLASTZ. Genome Research 2003, 13:112.