The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Using Simulated Annealing to Declutter Genome Visualizations

Jorge Núñez Siri,[1] Eric Neufeld,[1*] Isobel Parkin,[2] Andrew Sharpe[2]

[1]Department of Computer Science, [2]Agriculture and Agri-Food Canada, University of Saskatchewan, Saskatoon, SK CANADA
eric.neufeld@usask.ca

Abstract

AccuSyn is an interactive browser that visualizes conserved synteny relations (similar features) in genomes, giving biologists insights into the evolutionary history and functional relationships between genes. Even simple organisms have huge numbers of genomic features, and raw synteny plots present a daunting clutter of connections of which to make sense. Using a mixed initiative approach, AccuSyn integrates simulated annealing, a well-known metaheuristic for optimization problems, with human interventions to offer non-experts a way to automate decluttering, eliminating a tedious manual bottleneck in the discovery of syntenic information. AccuSyn has since been deployed online to a world user community.

Introduction

Recent studies show that due to innovations in DNA sequencing technology, data analysis is replacing data generation as the rate-limiting step in genomic studies (Neilson et al, 2010). AccuSyn partially automates the previously manual task of effectively visualizing of syntenic relationships within a genome or among different genomes.

Synteny refers to similar DNA sequences grouped into blocks. Whole chromosomes that match in size and shape are called homologous chromosomes, or homologs. Diploid organisms, like humans, have two homologous copies of each chromosome, one from each parent. Specifically, humans have a total of 46 chromosomes, 22 pairs of homologs and one pair of sex chromosomes (X and Y). Organisms containing more than two sets of homologs are called polyploid, a state especially common in plants, like Brassica napus, an amphidiploid or allotetraploid. Tetraploid means that it has four copies of each homolog or two diploid sets from each parent, and the prefixes *allo* or *auto* are used when these copies come from different or the same species, respectively. The B. napus genome has 19 chromosomes, derived from two diploid genomes, B. rapa with 10 chromosomes and B. oleracea with nine chromosomes. Therefore, B. napus is called allotetraploid because its first 10 chromosomes are homologs of B. rapa and its last nine chromosomes are homologs of B. oleracea (Cheng *et al*, 2014). Such clearly observable features provide evidence for major evolutionary events in the homology of species, defined as shared ancestry between two similar biological sequences (genes) (Tang *et al*, 2015).

Accusyn was created as part of the Plant Phenotyping and Imaging Research Centre (P^{2}IRC), managed by the Global Institute for Food Security at the University of Saskatchewan. P^{2}IRC uses the MCScanX software (Wang *et* al, 2012) to detect syntenic blocks using inputs from a Genomic Feature Format (GFF) file and a BLAST (Basic Local Alignment Search Tool) (Altschul *et al*, 2016) file, which contains sequences of interest. BLAST appears to be the most popular tool for finding similar (i.e. approximate) measures. Together these files make it possible to organize information about matching block pairs, including location and measurements of alignment strength (Wang *et* al, 2012).

MCScanX displays results using several plot types. We concentrated on circular plots, which arrange chromosomes in a circle and connect syntenic blocks with ribbons giving a compact overview of the genome's synteny.

Figure 1 depicts the decluttering and enhancement process for the recently sequenced Chinese Spring Wheat genome (IWCSG, 2018). The upper left shows the initial syntenic graph with just one ribbon color, which implies much synteny but does not offer much detail. The upper right image shows how simply coloring the blocks suggests the presence of homologs. Filtering ribbons by block length (middle left) further simplifies the view.

The middle right image shows an early result of our mixed initiative decluttering algorithm on filtered links, more strongly suggesting homologs. The bottom diagrams add back the filtered links and add other enhancements.

Figure 1. Top to bottom, left to right: All links one color; links colored to match sources with targets; filtered links, links grouped after decluttering, unfiltered presentation after groups, and final diagram with additional tracks of information.

Background

The first plant to be completely sequenced was Arabidopsis thaliana (AGI, 2000). Figure 2 shows compares syntenic connections for this five chromosome genome with that of Chinese Spring Wheat. Arabadopsis has 211 conserved blocks, each one connecting up to 188 pairs of genes. By comparison, Chinese Spring Wheat, a recently-sequenced plant genome (IWCSG, 2018), has 21 chromosomes for a total of 1,299 syntenic blocks, each connecting up to 2,253 pairs of genes. The Wheat genome (right) has three subgenomes A, B, and D, each with seven chromosomes, which permits other decluttering opportunities (Figure 5).

MizBee (Meyer *et al,* 2009), a stand-alone synteny browser and the first with multiple side-by-side linked views to show conserved syntenic relationships at different scales, provided much inspiration for this work.

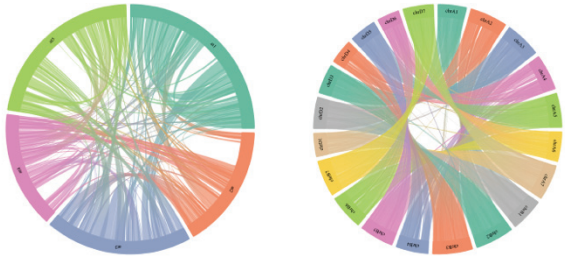

Figure 2. AccuSyn overview of syntenic links in Arabadopisis (left) and Chinese Spring Wheat (right).

The MizBee paper hints at the clutter problem in one of its case studies and tackles the problem by using edge bundling to join connections from neighboring blocks. For a discussion of this pioneering work and a survey of other synteny browers using circular plots, see (Nunez Siri, 2019), and for new work on other visualization styles, see (Bandi, 2019).

Our core challenge was reducing the time spent manually generating synteny diagrams by creating software accessible to non-experts. We began by looking for graph-theoretic methods of embedding graphs in the plane with a minimal crossing number. Although answering the crossing number decision problem is NP-complete (Garey and Johnson, 1983, Schaefer, 2013), we hoped to find good heuristics. However, our core data structure, which represents blocks as vertices and syntenic relations as edges, does not trivially map onto a typical graph as all blocks within a chromosome must move as a unit. We investigated a polynomial-time divide and conquer approximation algorithm (Shahrokhi *et* al, 2001) for the bipartite crossing number problem within a factor of $O(\log^2 n)$ from the optimal graph. However, given that syntenic graphs are not generally bipartite, we used simulated annealing (SA), a versatile tool for finding good solutions to combinatorial problems, which evolved into a "human-in-the-loop" solution for reasons that will become clear below.

Proposed Solution

Decluttering with Simulated Annealing

Here we sketch key features of simulated annealing, a search optimization algorithm that begins from an initial state that it perturbs, measuring the result to determine whether the new configuration has improved with respect to an objective function. If it has, it accepts the new configuration. If not, it accepts it with probability $p(t)$ where t is time and $p(t)$ decreases with time. This gives the algorithm an opportunity to try another configuration when it gets stuck in a local minimum. The algorithm then continues to run for either a finite number of iterations or time.

SA is a well-established technique (Russell and Norvig, 2009). The key for novices is casting the problem into a

form that the algorithm can use. For this work, the initial configuration was the order of the chromosomes arranged along the circumference, which was perturbed by swapping locations of random chromosomes. The objective function, the number of link crossings, was chosen after noticing that some syntenic maps could be drawn in in bands – making the graph resemble a basketball – or in clusters or bubbles (see Figure 5). These functions are simple but often yield marginal improvement: it was difficult sometimes to tell by inspection whether a small number of swaps resulted in a meaningful visual improvement, so the software printed the resultant number of crossings.

SA research has discovered a variety of patterns for constructing perturbation algorithms (annealing schedules) and objective functions. What was surprising and timely was how well the basic SA algorithm worked in this new domain in tandem with human interventions. Early on, we noticed that the algorithm performed well at the outset, but floundered mid-problem. We initially ran the algorithm repeatedly, using a previous solution as a starting point, but this exposes a weakness of our simplistic annealing schedule: the monotonically decreasing probabilistic threshold makes the algorithm adventurous initially, then increasingly conservative once it has invested a good deal of time into a "reasonably" good configuration. Running the algorithm twice rather than running it for a longer time amounts to momentarily throwing caution to the winds in hopes of a breakthrough and might make things much worse. Interestingly, (or fortuitously), at the point the algorithm thrashed, a good next move seemed intuitively obvious to the user, but we could not translate this intuition into the annealing schedule, so we added the human to the loop. The user could start by letting the software a few automated detanglings, step in with manual interventions when progress stalled, then restart the algorithm. This will be explained in detail in the section on *human-in-the-loop*. First, we will describe some challenges building the objective function.

Minimizing Crossings

In Circos, chromosomes are represented as arcs along a circle's circumference and syntenic connections between blocks on chromosomes are drawn as curves. Although strictly speaking, a chord is a *straight*-line segment joining two points on a circle, here we use the term *chord* rather than *curve* to be consistent with the Circos documentation.

Chords make it hard to calculate crossings exactly because the chord parameters needed for these calculations, are buried within Circos. To simplify this, AccuSyn projects each chord onto the straight-line segment connecting its endpoints, and assumes that if two chords intersect, their projection onto straight links will also intersect.

Figure 3 illustrates why this works well.

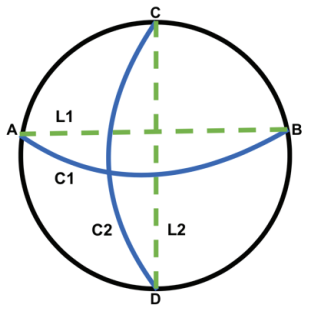

Figure 3. Chords C1, C2 and their projections L1 and L2

Let C1 and C2 be chords inside a circle, and L1 and L2 be their corresponding projections onto line segments. Consider a projection that maps C1 to L1. Because L1 divides the entire plane into two disjoint regions, each endpoint of C2 is in one of these two disjoint regions. Therefore, any line segment connecting them must cross L1. It may be that some edge cases of the library's internal chord drawing algorithm results in crossings not getting counted incorrectly, another good reason to have a human-in-the-loop approach.

This observation, and some basic computer graphics arithmetic offered a quick computation of edge crossings.

Human-in-the-loop

Because our SA algorithm was basic, it worked best at the outset, possibly because the number of crossings was huge, and many choices yielded an improvement. Later, it seemed to get "stuck". We conjecture that this was because it seemed that a specific sequence of choices was needed to break the log jam. Thus, we integrated manual movement of a chromosome by the user into the software, after which the software often achieved several successes. This approach to hard problems is sometimes called "human-in-the-loop", "cognitive orthoses" (Ford *et al*, 2015) or "mixed initiative" (Makonin *et al*, 2016).

A first step was to filter the connections to show blocks with "At Least" or "At Most" *x* connections, reducing the size and tractability of the problem, and presenting clearer intermediate visualizations.

Another useful intervention is flipping of chromosomes, that is, inverting the locations of the blocks inside a chromosome. Figure 4 compares two Wheat subgenomes before and after flipping chromosomes. Such crossings cannot be eliminated by moving chromosomes.

Our system also assists the human by allowing the storage of intermediate configurations as thumbnails to return to, a cognitive orthosis approach to "intelligent backtracking".

Figure 5 shows two diagrams produced using this approach. The main features of each diagram were produced in less than fifteen minutes.

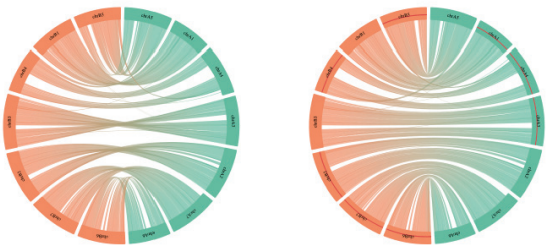

Figure 4. Before flipping (left) and after.

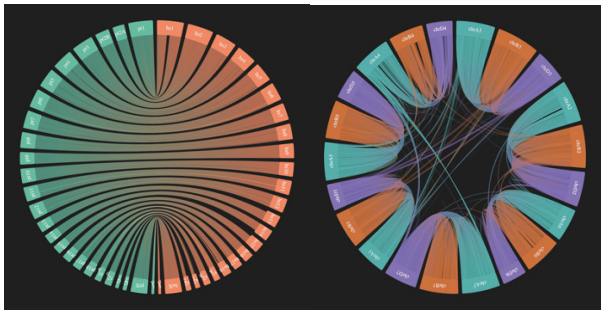

Figure 5. A basketball view comparing humans and chimpanzees, and a bubble view on the triplicated Wheat genome.

Summary and Discussion

Our interdisciplinary team was challenged with the problem of automatically decluttering synteny diagrams, a current bottleneck for practitioners. Initial successes were obtained by casting the problem abstractly, using SA to declutter the diagram to a point where it was useful allow human interventions. This first cut of the software, deployed at https://accusyn.usask.ca, has been well-received by its user community, accomplishing a significant and potentially ongoing lateral transfer of knowledge.

A keen and careful referee pointed us to several relevant references on SA, notably the Modified Lam schedule (Lam and Delosme, 1988) which we expect will help address the thrashing problems as project enters its next phase.

Acknowledgements

Thanks to P^{2}IRC and the University of Saskatchewan for support of this research and to the anonymous reviewers.

References

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796-815.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403-410.

Bandi, V. 2019. SynVisio: Synteny Browser. https://synvisio.usask.ca, 2018. Human-Computer Interaction Lab. University of Saskatchewan. Retrieved June 11, 2019.

Cheng, F., Wu, J., and Wang, X. 2014. Genome triplication drove the diversification of Brassica plants. *Horticulture Research* 1:14024.

Ford K., Hayes, P., Glymour, C., and Allen, J. 2015. Cognitive Orthoses: Toward Human-Centered AI. *AI Magazine* 36:5–8.

Garey, M. and Johnson, D. 1983. Crossing number is NP-complete. *SIAM Journal on Algebraic Discrete Methods* 4(3):312-316.

The International Wheat Genome Sequencing Consortium (IWGSC). 1028. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361(6403).

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* 19(9):1639-1645.

Lam, J., Delosme, J. 1988. Performance of a new annealing schedule. In *Proceedings of the 25th ACM/IEEE Design Automation Conference* 306-311.

Makonin, S., McVeigh, D., Stuerzlinger, W., Tran, K. and Popowich, F. 2016. Mixed-Initiative for Big Data: The Intersection of Human + Visual Analytics + Prediction. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* 1427-1436.

M. Meyer, T. Munzner, and H. Pfister. MizBee: A Multiscale Synteny Browser. 2009 *IEEE Transactions on Visualization and Computer Graphics* 15(6):897-904.

Nielsen, C.B., Cantor, M,. Dubchak, I. Gordon, D., and Wang, T. 2010. Visualizing genomes: techniques and challenges. *Nature Methods*, 7(3s):S5-S15.

Núñez Siri, J. 2019. *AccuSyn: Using Simulated Annealing to Decultter Genome Visualizations*, M.Sc. thesis, University of Saskatchewan, Department of Computer Science.

Russell S., and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*, 3rd edition, Upper Saddle River, NJ: Prentice Hall.

Schaefer, M. 2013. The Graph Crossing Number and its Variants: A Survey. *The Electronic Journal of Combinatorics*, 1000:21-22.

Tang, H., Bomho, M.D., Briones, E., Zhang, L., Schnable, J.C., and Lyons.E. 2015. SynFind: Compiling Syntenic Regions across Any Set of Genomes on Demand. *Genome Biology and Evolution* 7(12):3286-3298.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J.C. and Paterson, A.H. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 40(7) e49.