Final Project:

Creating a Synteny Map Alignment Between the Genomes of Two Species

MCDB 752: Genomics & Bioinformatics

December 13, 2006

Elizabeth Wojcik

I. Introduction: What is Synteny and Its Importance in Biology?

The goal of synteny mapping is to align chromosomes or genomes between two or more species. Aligning genomes to each other will allow for the identification of homologous regions. With the belief that homologous sequences share similar biological function(s), identifying what a gene or region of sequence does in one organism will theoretically shed light onto what another homologous gene/sequence does in a second organism. This belief is firmly planted in the scientific community and is shown, in one instance, with the use of mice as model organisms for the study of human diseases.

II. General Concerns with Alignments

The first concern when creating an alignment is to determine what type of dataset to align. While the goal of synteny mapping in this paper's context clearly defines the dataset as the nucleotide sequence with a length of the whole genome, protein sequences may also be used in alignments and thus should be considered under the general concerns of creating an alignment. DNA sequences contain four nucleotides, A, G, C, and T that occur throughout the sequence. In the open reading frame (ORF) of a gene, a mutation in one of these nucleotides at a particular site will not necessarily change the encoded protein sequence due to the third base wobble allowance in the codon.[11] Because of this allowance of mutations without consequence, a high probability of further sequence divergence exists. Other regions such as promoter regions or genes that encode RNA molecules may not be as forgiving, but can occasionally accept mutations. For protein sequences, changes in the sequence is thought to have a higher probability of having a negative effect on the individual since there is no wobble effect like in the ORFs. Due to this theory, protein sequences may then be considered to be more conserved throughout evolutionary history. However alignments using only protein sequences would miss taking into

account all other regions in the genome which may provide insight into the homology between the two species.

From a biological view, comparing DNA sequences is slightly more complicated than comparing protein sequences since DNA is double stranded. For example if you have two species each with the following identical sequences:

Species 1: *Top Strand* 5'ATGCTTGG3' Species 2: *Top Strand* 5'CCAAGCAT3'

Bottom Strand 3'TACGAACC5' *Bottom Strand* 3'GGTTCGTA5'

If the query was the top strand of Species 1 and the search was running against the top strand of Species 2, the two sequences would not align. However if the top strand of Species 1 was run against the bottom strand of Species 2, an alignment could be found as long as the program can look for the word in the reverse orientation. This whole problem is due to the fact that sequences can sometimes translocate or invert themselves into a chromosome by a variety of methods.[5, 6] So even if a gene had one orientation in a common species at one time, when that species diverged into two species, the orientation within the DNA strands could change in either of the new species. In a protein alignment, only one strand exists so it is easier to align.

Duplication events also complicate matters in an alignment. For instance a species has gene X and then this species diverges into two species (Species A and Species B), each with the gene X. If Species A has a duplication event of gene X, one copy (gene X1) will retain the original function while the new copy (gene X2) is relatively free to mutate. Even if the new copy (X2) is mutated so that it now has a new function, Species A can most likely still survive because the original copy (X1) is still performing its original function. This however is difficult for alignments. Even though both copies of the gene (X1 and X2) originated from the same gene X, they will not have the same similarity scoring when compared to Species B gene X. This makes

it hard to determine that both gene X1 and X2 came from the same gene X and at one point in evolution had the same function.

Another concern when creating an alignment is determining the scope of an alignment which is does by asking what the purpose of the alignment is in the first place. When comparing two nearly identical sequences of the same length, alignment by either local or global alignment methods should produce the same results.[12] Local alignment focuses on small regions of sequences and tries to arrange two or more sequences together. While this method works well when comparing domains or motifs in sequences, it may not align all the sequences in a genome. This method does not force alignments to happen, but rather aligns the best homologous regions. Global alignment on the other hand seeks to align all regions in the genome. While small areas of regions might not have the best fit (as when using local alignment), the overall best fit is found. This may provide a broader view as to how two or more sequences are related and is thus more useful when comparing two very distantly related species. In synteny mapping, global alignment should be used to accomplish aligning every area of the genome of a first species to every area of another species' genome.

A final general concern discussed in this paper arises from the difficulties of creating global alignments when comparing the genomes of two species that are different lengths. For instance if the human genome (3.2 billion basepairs) was compared to the laboratory strain *E. coli* K12 genome (4.64 Mb) [3], an almost certain probability exists that a large number of regions in the human genome will not align with any decent scoring to any region of the *E. coli* genome. However, the *E. coli* genome has a higher probability of aligning all of its regions of DNA with the human genome due to the bacteria's smaller size. So the size of the two genomes as a general

parameter of successfully aligning all regions of the genomes must be taken into account when doing global alignments.

III. Methods to Align Sequences:

Many methods have been employed over the years in order to align sequences from two different species. Before the use of computers, the first method that was done by eye is to look for identical sequences using the first letter in a query sequence, going through the genome to locate where that letter is present, and then making a list of those locations. From here, the second letter in a query is looked at in the first letter's list, and so on. While this is a systematic method, it will not work for instance in finding a homologous region in the genome if that region is missing the first letter of the query's sequence.[7] Also, this method does not take into account substitutions, insertions, or deletions.

Another method is called the Dot-Matrix and is more of a global alignment than a local alignment. On a square graph, one sequence is placed on the x-axis while the other sequence is placed on the y-axis. A dot is positioned in the box on the graph when the row and column shared by that box has the same letter (or nucleotide in our case).[3] From this general method, similarities in a sequence can be identified. However, quantification of the areas that are not identical is lacking and thus new methods were created.

Needleman and Wunsch in 1970 came out with a dynamic programming method which finds an optimal global alignment. While they originally created it for protein sequence alignment, this method can be applicable to nucleotide sequences. A matrix similar to the dot matrix is formed but instead of a dot going into a box, the amount of 1 is placed. Starting from the bottom right of the matrix, successive summations are done row-by-row to create a path through the matrix showing the maximum match contributors of the aligned sequences. Penalty

values were given in a box when a gap in the sequence was made. Similarity values that could be incorporated into the matrix in the occurrence of a mismatch were discussed.[8] While this model is a global alignment which is what is required in a synteny mapping alignment, one main drawback exists. Since a whole matrix is computed in this method and synteny mapping wants to align two huge genomes, the time necessary to complete the matrix is great. If the time taken to create the matrix was decreased or if the calculations of the matrix ceased after a certain threshold value, the Needleman-Wunsch could be used for synteny mapping.

A second algorithm, similar to the Needleman-Wunsch algorithm, was created in 1981 by Smith and Waterman.[10] This dynamic program looks to find the optimal local alignment, unlike the Neeldeman-Wunsch algorithm that finds the optimal global alignment. One similarity is that the time and memory needed to create an alignment is proportional to the lengths of the compared sequences.[7] In other words, it would take too long to compare two genomes, is only a local alignment, and as a result is not a correct choice for synteny mapping.

A potentially faster method that will improve finding homologous regions in a genome is to search using 'words' from the query. This method was used in creating the first BLAST program in which short sequence words (usually a few letters) are looked for in the entire length of the compared sequence.[1] If a word is found, this is called a 'hit'. From here, extensions would take place in which the computer looks beyond the few letters that make up the word and sees if the surrounding sequence matches the query. While this program will pullout identical regions to short queries (a.k.a.local alignments), for the scope of aligning genomes and taking into account other factors such as gaps, BLAST will not suit synteny mapping needs.

A second version of BLAST called gapped BLAST improves the search through the data by searching with two non-overlapping words from the query. The computer looks for regions in

the sequence that has two hits with a distance of "A" (A is equal to the distance between the two words in the query) and then invokes an extension only if this parameter is met.[2] This allows for faster searching due to the decreased amount of extensions initiated. Gapped BLAST also allows for gaps in the alignment of the sequences either from insertions or deletions, but is still considered a local alignment.

FASTA is similar to the Gapped BLAST method because it initially looks for multiple words in the sequence. After this, it scans these hits using a scoring matrix and keeps the best scoring regions. Next, regions with scores above a threshold value determined from a similarity matrix are joined together with "joining penalties" taken into account. Finally, a new optimal alignment is determined with taking into account the highest scoring initial regions. [9] FASTA is at first a local alignment in that it seeks local similarities between the two sequences, but then through the last step tries to become a global alignment when it forces regions between two best scoring regions to align. However, since FASTA is from a modified version of the Smith and Waterman algorithm, it is not applicable to the scale of comparison needed for synteny mapping. The Smith and Waterman algorithm gets too complex and time consuming when using sequence sizes over one hundred thousand base pairs.[4]

IV. Technique to Create a Synteny Map Between Two Species

In the end, it is difficult to create a perfect, optimal synteny mapping alignment between two species. The idea of global vs. local alignment must be decided. Since the definition of synteny in the context of this paper is to create an alignment between two genomes, a global alignment is ideally wanted. A Dot Matrix or Needleman-Wunsch can provide a global alignment. However, the Dot Matrix does not take into account gap values or any values at all. It only clearly identifies the location of exact matches of sequences. Needleman-Wunsch takes into

account gap values and substitution values, but for the size of comparing two large genomes would take an enormous amount of time to create the alignment. Smith and Waterman will only cover taking into account gap penalties and substitution values, but is again size limited and also is a local alignment program.

The size of two sequences is a severely limiting factor when making a synteny map. In order to cut down on the time, the first species' genome could be chopped into section either by individual chromosome or regions on a chromosome. These pieces could then be aligned to the second species genome, on at a time. Two main drawbacks to this method exist: 1. All the regions of the second species might not get aligned to the first genome, so this method would have to be done where the second species genome is cut into pieces and then aligned to the first genome. This is a problem related to the genome size concern discussed in section II of this paper. Drawback two is to determine where to make a cut in the genome and/or chromosome. Because of the general large size of chromosomes, a cut must be made somewhere in a chromosome instead of making the pieces for the alignment the individual whole chromosomes. If that cut is made in an ORF, the alignment of that gene will not be optimal. If the cut is made in a non-coding region, that region could be conserved and because of the cut not show up in the alignment as a high scoring match. In order to overcome this, an additional alignment series must be done with the cuts made in different regions compared to the initial alignment.

This method of cutting a genome into manageable pieces should still be considered a global alignment, even though it slightly resembles local alignments in that pieces of the genome are looked at instead of usually the whole genome at once. However, because of two reasons I still consider this method to be a global alignment: First, the size of the piece will be larger than that of a normal query local alignment search. Second, all the pieces of one species' genome will

be looked at in its entirety and aligned to the second species genome. The reverse for the second species' will be done so that all the pieces of the second genome is aligned to the first genome. This satisfies the main part of the general goal of a global alignment in which all of both genomes must be aligned to one another.

This decrease in time would entice the use of the Needleman-Wunsch algorithm (remembering that the time and memory it takes to produce this alignment is proportional to the lengths of the sequences[7]). However, the mathematical completion of the whole scoring matrix even in regions away from the optimal maximum match takes up a good amount of time. If a limit of what is looked at could be set like in a gapped BLAST, that would decrease the time needed to run the program. Determining the initial area or seed for this focus is a problem. Focusing on that single seed is why gapped BLAST is considered a local alignment; it does not look at the whole picture only the highest scoring piece of the sequence. So decreasing the time it takes by decreasing the area looked through cannot be done unless the global scale of the alignment is sacrificed.

The creating of a synteny mapping alignment is definitely a difficult endeavor because of the caveats of time/memory needed to run a program and keeping a global focus instead of a local view. Taking into account the modifications mentioned above in section IV and also paying attention to the general concerns mentioned in section II, a modified version of the Needleman-Wunsch could be used to create a synteny mapping alignment. While it would not be perfect in the sense that time would most likely still be a limiting factor as to the size of genomes compared, it is the best option compared to the other methods discussed in the paper.

References:

1. Altschul, S.F. et al. "Basic Local Alignment Search Tool." J Mol Biol. (1990) Oct 5;215(3):403-10.

2. Altschul. S.F. et al. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." Nucleic Acids Research. (1997) 25 (17): 3389-3402.

3. Brown, T. A. Genomes. 2nd Ed. Oxford: BIOS Scientific Publishers Ltd, (2002).

4. Dubchak, I. and Pachter, L. "The Computational Challenges of Applying Comparative-Based Computational Methods to Whole Genomes." Briefings in Bioinformatics. (2002) 3 (1):18.

5. Gordon, A.J.E. and Halliday, J.A. "Inversions with Deletions and Duplications." Genetics. 1995 May; 140(1):411-414.

6. Griffiths, A.J. et al. An Introduction to Genetic Analysis. 7th ed. New York: W.H. Freeman & Co., (1999).

7. Koonin, E.V. and Galperin, M. Y. Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. Norwell: Kluwer Academic Publishers, 2003.

8. Needleman, S. B. and Wunsch, C. D. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins". J. Mol. Biol. (1970) 48, 443-453.

9. Pearson, W.R. and Lipman, D.J. "Improved Tools for Biological Sequence Comparison." Proc. Natl. Acad. Sci. (1988) April. 85: 2444-2448.

10. Smith, T. F. and Waterman M. S. "Identification of Common Molecular Subsequences." J. Mol. Biol. (1981) 147; 195-197.

11. Strachan, T. & Read, A.P. Human Molecular Genetics 2. Oxford: BIOS Scientific Publishers Ltd., (1999).

12. Wikipedia. "Sequence Alignment" Wikipedia Foundation, Inc. Last modified: December 7, 2006. Accessed December 11, 2006.