Synteny – Finding meaning in order

Emmett Sprecher
December 13, 2006
Genomics & Bioinformatics (CBB 752)

The concept of measuring sequence similarity is not a novel one in biological investigation. From DNA to proteins, measuring the similarity of sequences is a useful tool for conveying information obtained about one sequence to the other, either by their similarities or differences. In the realm of genomics, the comparison is on a much larger scale. While still analyzing DNA, comparisons can be done between the organizations of larger order genetic components in two organisms' genomes or chromosomes. Such organizations of genetic material on a species' chromosomes are termed synteny maps, as they give the relative positions of all the syntenic genes (two genes are said to be syntenic if they reside on the same chromosome).

Synteny maps can convey several types of information. For one, they can give a larger scale perspective on whether two genomes are evolutionarily, closely related. Furthermore, syntenic maps can give an evolutionary perspective on how genomes have changed and diverged over time (Delseny; Bennetzen and Freeling; Bellgard et. al.). The chromosomal rearrangements and genetic sequence modifications that aggregate over the process of speciation cause the differentiation in synteny maps, therefore they are a large scale look at the pattern of genetic divergence. Even very closely related genomes that were thought to be highly syntenic have greater degrees of differentiation than originally believed (Delseny; Feuillet; Schmidt).

Both the divergence and the similarity can have further functions. Based on information known about one of the regions, inferences can be made about the other (Delseny; Bellgard et. al.; Schmidt). For example, it is possible to predict the location of yet undiscovered genes due to their presence in a characterized region that is syntenic to an uncharacterized region in another genome (Lindqvist et al). Even more interestingly, it is sometimes possible to infer genes from differences between syntenic regions. An example here is the case of resistance genes in certain cereals. Due to the high mutation rates associated with resistance genes, they were discovered to be places of divergence in otherwise syntenic regions (Bellgard et. al.; Schmidt). Beyond specific gene location, the similarities and differences can lend researchers perspective on the associated phenotypic similarities and differences between related species, and also suggest insights into what differentiates species.

Synteny mapping, as indicated above, are mappings of larger gene segments, which are often done by aligning genes and genetic markers rather than individual nucleotides. In this way it is possible to look at larger regions of significance quickly and globally. Markers such as expressed sequence tags (ESTs) and restriction fragment length polymorphism (RFLP) are commonly used, while the actual investigation of homologous genes is often only done on further investigation (Delseny; Schmidt). This gives an easier and systematic way of aligning regions without having to do exact sequence comparisons, although it has its own associated difficulties. Exact sequence comparisons are done as well to gain additional perspective, but they are subject to the same problems as local sequence alignments, except multiplied over larger regions, and in cases where rearrangements are expected. Of far greater significance, doing exact

sequence comparisons between chromosomes or large DNA segments requires that those sequences actually exist. And even if they do exist, the organisms' sequences can have different levels of detail and accuracy. Despite these limitations to direct sequence comparison, it is still one of the more potent tools and is often utilized when exact sequence data is available (Delseny).

Another of the primary methods of mapping syntenic regions is through the use of homologous probes that cross-hybridize between the species being compared (Delseny; Schmidt). The probes are generated from gene products (mRNA) and can thereby hybridize to the original genes in both genomes as long as they both have the same original gene.

But with all of these approaches, problems remain.

For cross-hybridizing probes, the selection of probes limits the loci that are detected. First by the nature of having to generate the probes, it must be for known genetic material in a region. Also, the degree of similarity between even related genetic material, will impact the hybridization. For example, if two homologous genes have diverged enough in the particular region that the probe binds, then the probe may not show on one of the genomes even though the genetic material is arguably still orthologous. This can also be a problem when using probes that have proven to work between two related species, then trying to use them on a third related species. Though one might think that the probes would work given the previously demonstrated commonality, the functionality of any probes depend completely on the specific organisms in question (Delseny).

The opposite problem is also an issue. Since many genes are members of gene families, if a probe is selected for an area of commonality between the genes in the family, it may bind indiscriminately to the gene family members. While this would provide some information, the confusion of homologues and paralogues decreases the informational granularity of the synteny maps. Therefore, when using cross-hybridizing probes, probe selection is of tremendous importance (Delseny).

On the related topic of ESTs, there are problems as well. ESTs are based on the mRNA products of transcription and splicing. Therefore, they are highly specific and may not fully correctly hybridize with related genes, even those that share original genes, particularly in the face of multiple splicing potentials. They are however useful in tracking the placement of multiple gene products that originate from the different splices of the same gene (Mayer).

RFLP analysis is subject to issues as well (Schmidt). In RFLP analysis, fragments are generated by treatment of a genetic sample with a restriction enzyme or set of restriction enzymes, and thereby generate a unique "fingerprint" based on the complex combinatoric pattern of the restriction sites. If the patterns are the same, then you can gain significant information about the synteny of the two chromosomes or genomes. If the RFLP patterns differ, however, the results are much harder to interpret. Whether due to elongation or shortening between fragments, or much worse, non-conservation of restriction sites, the changes can compound to form unpredictable results as you get further away from exact conservation of sequence.

Another set of issues arises when resorting to actual, direct sequence comparison, and that has to do with the reliability of the sequences being compared. While the great

power of and interest in comparative genomics comes from the increasing availability of greater numbers of closely related genomes on which to compare, the sequences are constantly being improved (Bellgard et. al.). In addition to new genomes, new drafts of published genomes are coming out on a regular basis. The push to get genome data out in the first place can require a sacrifice in accuracy of the initially published sequences with later attempts made to increase the precision (Roberts). But even more so, there are some areas that are tremendously difficult to sequence (such as highly repetitive sequences, which in and of themselves make up a decent portion of syntenic differences between grass species).

So, how can we address the problems of synteny mapping? How can we increase our detection sensitivity for creating synteny maps? The fact that the markers and genetic sequence searches are done separately, and considerations of micro- and macro-synteny are performed for the same genome comparisons seems to indicate that there might be benefits to integrated approaches for synteny detection (though some algorithms already exist (Liao et. al.; Vandepoele et. al.)). Certainly there is room for improvement by utilizing the strengths of each sequence comparison, feature based analysis (as suggested in Bellgrad et. al.) and some of the marker and RFLP comparisons.

However, the corollary question may perhaps be more important and interesting: how can we increase the usefulness of synteny maps once we create them? The actual generation of synteny maps will likely improve with time, improved genome datasets, and effort towards integrative algorithmic developments, but what about our

understanding of the intrinsic impacts of structural arrangement on gene expression and resultant phenotypes?

This is really the ultimate goal of genetic science. The complete understanding of how we go from molecules to living organisms, so it is of paramount importance in that sense. And while I am not suggesting an immediate leap to complete understanding of the relationship between structural arrangement and gene expression influences, any improvements in this area feed back onto synteny mappings, allowing the knowledge gained to be applied all syntenic regions. At the same time, the increase in sequenced and tagged, related organism genomes, particularly closely related ones, increasingly allows for the derivation of these spatially related influences (Bennetzen and Freeling). It is a feedback loop. The more we know, the more we can derive. Yet spatial influences have remained beyond our reach for so long due to their interactive complexity, that perhaps we have forgotten to reconsider the openings we have with our increased available knowledge. As suggested at the end of the Bellgard et. al. review, perhaps it is time to consider this area again.

Therefore, the near future should consist not only of improvements in the generation of syntenic maps, but also in the investigation of complex, higher-order, genomic interactions. Investigations could utilize the increasingly accurate and expansive genomes available of closely related organisms, whose similarities and differences can help to highlight import effects. But alternatively (and in parallel) creation of more controlled experiments by creating particular genome modification events in existing organisms and comparing them to the original species could help yield information about the direct impacts of particular deviations.

Therefore I would propose a two-pronged approach towards future investigations into synteny mapping and its applications. First, I would suggest an integrative approach towards the generation of tools for detecting synteny. Given the complexity of the problem, and the implications of the fact that multiple layers of analysis are done on the same genomes in attempts to parse out more information. As the common search formats have proved for basic sequence comparisons such as BLAST and FASTA, the tools that permit quality, fast result findings will in and of themselves aid in driving research in the area.

The second prong of the approach would be as suggested above: to begin concerted efforts toward simplified experiments for distilling the rules governing spatial effects in genomic arrangement.

Such experiments would focus on the fundamental rearrangements that can impact synteny, such as chromosomal recombination, genetic inversions, changing of the distance between genes (either by extension or contraction of the intergenic repeats), and transpostion (Bennetzen and Ma, Bellgard et al). For example, one could take the rice genome and in well-studied segments, do comparative studies on the impacts of reversing given genes, or switching their strand. By using a single species rather than using the closely related other cereal genomes, the results should be that much easier to distill. The advantage in such a situation is that the experiment eliminates most variables. The difference in genetic expression, if any, controlled for by the inter-organism deviations between unmodified rice plants, should be largely due to the modification made. Additionally, having the close wheat, maize, and other genomes, allows quick extensions

of any findings to the other cereals to see if they are consistent. In fact, known inversions between rice and wheat could be the basis for selecting target genes to experiment on.

Through this hybrid approach we are able to simplify the problem of deriving higher order structural impacts while still utilizing the increasing available genomic information. Simplification is a necessary step because there are still considerable complexities in comparing even apparently very similar genomes (Bennetzen and Ma). But the fruits of these simplifications would have the potential to fill in the gaps between our current understanding, and the next step up of understanding deviations between closely related species.

So while these only characterize a path towards a small connective step, the first step is always the hardest. And more importantly, it brings us one step closer to a more comprehensive understanding of genomics and our biological basis.

References:

Bellgard M, Ye J, Gojobori T, Appels R: **The bioinformatics challenges in comparative analysis of cereal genomes – an overview**. *Funct Integr Genomics* 2004, **4**:1-11.

Bennetzen JL, Freeling M: **The Unified Grass Genome: Synergy in Synteny**. *Genome Research.* 1997 **7**: 301-306.

Bennetzen LJ, Ma J: **The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis**. *Current Opinion in Plant Biology* 2003, **6**:128–133.

Delseny M: **Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement**. *Current Opinion in Plant Biology* 2004, **7**:126-131.

Feuillet C, Keller B, **Comparative Genomics in the Grass Family: Molecular Characterization of Grass Genome structure and Evolution**. *Annals of Botony* 2002, **89**: 3-10.

Liao BY, Chang YJ, Ho JM, Hwang MJ: **The UniMarker (UM) method for synteny mapping of large genomes**. *Bioinformatics* 2004, **20** 17: 3156–3165.

Lindqvist AKB, Alarco´n-Riquelme ME: **The Genetics of Systemic Lupus Erythematosus**. *Scand. J. Immuno.* 1999, **50**:562–571.

Mayer K and Mewes H-W, **How can we deliver the large plant genomes? Strategies and perspectives**. *Current Opinion in Plant Biology* 2001, **5**:173–177.

Roberts L.: **Controversial From the Start**. *Science* 2001, **291**: 1182-1188.

Schmidt R, **Synteny: recent advances and future prospects**. *Current Opinion in Plant Biology* 2000, **3**:97–102.

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y: **The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between Arabidopsis and Rice**. *Genome Research* 2002, **12**: 1792-1801