Comparative genomics research benefits

ASSIGN BUSTER

ABSTRACT

The rapidly emerging field of comparative genomics has yielded dramatic results. Comparative genome analysis has become feasible with the availability of a number of completely sequenced genomes. Comparison of complete genomes between organisms allow for global views on genome evolution and the availability of many completely sequenced genomes increases the predictive power in deciphering the hidden information in genome design, function and evolution. Thus, comparison of human genes with genes from other genomes in a genomic landscape could help assign novel functions for un-annotated genes. Here, we discuss the recently used techniques for comparative genomics and their derived inferences in genome biology.

INTRODUCTION

Comparative genomics is the study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. While it is still a young field, it holds great promise to yield insights into many aspects of the evolution of modern species. The sheer amount of information contained in modern genomes (750 megabytes in the case of humans) necessitates that the methods of comparative genomics are automated. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome.

Human FOXP2 gene and evolutionary conservation is shown in a multiple alignment (at bottom of figure) in this image from the UCSC Genome Browser. Note that conservation tends to cluster around coding regions (exons).

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral).

SCOPE OF COMPARATIVE GENOMICS

One of the important goals of the field is the identification of the mechanisms of eukaryotic genome evolution. It is however often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. For this reason comparative genomics studies of small model organisms (for example yeast) are of great importance to advance our understanding of general mechanisms of evolution.

Having come a long way from its initial use of finding functional proteins, comparative genomics is now concentrating on finding regulatory regions and siRNA molecules. Recently, it has been discovered that distantly related

species often share long conserved stretches of DNA that do not appear to code for any protein. One such ultra-conserved region, that was stable from chicken to chimp has undergone a sudden burst of change in the human lineage, and is found to be active in the developing brain of the human embryo.

Computational approaches to genome comparison have recently become a common research topic in computer science. A public collection of case studies and demonstrations is growing, ranging from whole genome comparisons to gene expression analysis. This has increased the introduction of different ideas, including concepts from systems and control, information theory, strings analysis and data mining. It is anticipated that computational approaches will become and remain a standard topic for research and teaching, while multiple courses will begin training students to be fluent in both topic.

Chromosomes from two genomes are drawn: human chromosome 1 (drawn with a global zoom factor of 50x) and mouse chromosomes 1-19, X, and Y with mouse chromosome 3 drawn enlarged 10-fold. Syntenic regions between human chromosome 1 and the mouse genome are connected by coloured curves (A), whose geometry and properties can be adjusted dynamically. Thus, all syntenic relationships with mouse chromosome 4 are coloured in orange (B), and all relationships falling within the 80-90 Mb region on human chromosome 1 are coloured in blue (C). Other relationships with alignments larger than 5 kb are coloured dark in grey (D) and all others are shown in light grey. The lines are drawn layered with light grey lines below all others, then dark grey, then blue and then orange. Although

approximately 44, 000 syntenic relationships are drawn, the use of a selective colour scheme maintains legibility. The outer track (E) is a histogram of the log density of syntenic regions over 100 kb windows on human chromosome

- GENOMES ARE MADE OF

Although living creatures look and behave in many different ways, all of their genomes consist of DNA, the chemical chain that makes up the genes that code for thousands of different kinds of proteins. Precisely which protein is produced by a given gene is determined by the sequence in which four chemical building blocks – adenine (A), thymine (T), cytosine (C) and guanine (G) – are laid out along DNA’s double-helix structure

BENEFITS OF COMPARATIVE GENOMICS

Using computer-based analysis to zero in on the genomic features that have been preserved in multiple organisms over millions of years, researchers will be able to pinpoint the signals that control gene function, which in turn should translate into innovative approaches for treating human disease and improving human health.

In addition to its implications for human health and well-being, comparative genomics may benefit the animal world as well. As sequencing technology grows easier and less expensive, it will likely find wide applications in agriculture, biotechnology and zoology as a tool to tease apart the often-subtle differences among animal species. Such efforts might also possibly lead to the rearrangement of our understanding of some branches on the

evolutionary tree, as well as point to new strategies for conserving rare and endangered species.

Comparative Genomics Goals

- Complete the sequence of the roundworm C. elegans genome by 1998.
- Complete the sequence of the fruitfly Drosophila genome by 2002.
- Develop an integrated physical and genetic map for the mouse, generate additional mouse cDNA resources, and complete the sequence of the mouse genome by 2008.
- Identify other useful model organisms and support appropriate genomic studies.

METHODLOGY

Genome correspondence

Genome correspondence, the method of determining the correct correspondence of chromosomal segments and functional elements across the species compared is the first step in comparative genomics. This involves determining orthologous (genes diverged after a speciation event) segments of DNA that descend from the same region in the common ancestor of the species compared, and paralogous (genes diverged after a duplication event) regions that arose by duplication events prior to the divergence of the species compared. The mapping of regions across two genomes can be one-to-one in absence of duplication events; one-tomany if a region has undergone duplication or loss in one of the species, or many-to-many if duplication/loss has occurred in both lineages. Fitch et al., developed a method called BBH (Best Bidirectional Hits), which identifies gene pairs that

are best matches of each other as orthologous. Tatusov et al., further enhanced this method, which matches groups of genes to groups of genes.

Understanding the ancestry of the functional elements compared is central to our understanding and applications of genome comparison. Most comparative methods have focused on one-to-one orthologous regions, but it is equally important to recognize which segments have undergone duplication events, and which segments were lost since the divergence of the species. Comparing segments that arose before the divergence of the species may result in the wrong interpretations of sequence conservation and divergence. Further, in the presence of gene duplication, some of the evolutionary constraints that a region is under are relieved, and uniform models of evolution no longer capture the underlying selection for these sites. Thus, our methods for determining gene correspondence should account for duplication and loss events, and ensure that the segments we compare are orthologous

Applications

Gene identification

Once genome correspondence is established, comparative genomics can aid gene identification. Comparative genomics can recognize real genes based on their patterns of nucleotide conservation across evolutionary time. With the availability of genome-wide alignments across the genomes compared, the different ways by which sequences change in known genes and in intergenic regions can be analyzed. The alignments of known genes will reveal the conservation of the reading frame of protein translation.

The genome of a species encodes genes and other functional elements, interspersed with non-functional nucleotides in a single uninterrupted string of DNA. Recognizing protein-coding genes typically relies on finding stretches of nucleotides free of stop codons (called Open Reading Frames, or ORFs) that are too long to have likely occurred by chance. Since stop codons occur at a frequency of roughly 1 in 20 in random sequence, ORFs of at least 60 amino acids will occur frequently by chance (5% under a simple Poisson model), and even ORFs of 150 amino acids will appear by chance in a large genome (0. 05%). This poses a huge challenge for higher eukaryotes in which genes are typically broken into many, small exons (on average 125 nucleotides long for internal exons) in mammals. The basic problem is distinguishing real genes – those ORFs encoding a translated protein product – from spurious ORFs – the remaining ORFs whose presence is simply due to chance. In mammalian genomes, estimates of hypothetical genes have ranged from 28, 000 to more than 120, 000 genes. The internal coding exons were easily identified using Comparative analysis of human genome with mouse genome.

Regulatory motif discovery

Regulatory motifs are short DNA sequences about 6 to 15bp long that are used to control the expression of genes, dictating the conditions under which a gene will be turned on or off. Each motif is typically recognized by a specific DNA-binding protein called a transcription factor (TF). A transcription factor binds precise sites in the promoter region of target genes in a sequence-specific way, but this contact can tolerate some degree of sequence variation. Thus, different binding sites may contain slight

variations of the same underlying motif, and the definition of a regulatory motif should capture these variations while remaining as specific as possible. Comparative genomics provides a powerful way to distinguish regulatory motifs from non-functional patterns based on their conservation. One such example is the identification of TF DNA-binding motif using comparative genomics and denovo motif. The regulatory motifs of the Human Promoters were identified by comparison with other mammals. Yet another important finding is the gene and regulatory element by comparison of yeast species.

Applications of comparative genomics to wheat

A number of important major traits requiring elucidation in wheat are essentially non-polymorphic. Thus there is no prospect of creating a mapping population which is the starting point of all positional cloning strategies in most species to date. Moreover given the size of the wheat genome, many traits lie in regions where the gene density per BAC is one or two, making it difficult if not impossible to walk from one wheat BAC to the next. The Ph1 locus (controlling chromosome pairing in wheat) is one such example, in which the starting point was wild type wheat and a mutant carrying a deletion of more than 70Mb (almost the size of the whole Arabidopsis genome). Its phenotype is not easy to score. My group wished to characterise this locus. We created three different types of mutagenised populations, sequenced the equivalent rice Ph1 region, built BAC libraries (all are now available free of IP) for Brachypodium (a small genome species more closely related to wheat), sequenced Brachypodium Ph1 equivalent region, built a hexaploid (CS) (737, 000 clones) wheat in collaboration with INRA (providing a further 500, 000 clones), exploited Jorge Dubcovsky’s

Tetraploid wheat BAC library, sequenced wheat BACs and defined the tissues in which the Ph1 phenotype is expressed. I will discuss the approaches adopted and resources created.

Application of comparative genomics to the analysis of vertebrate regulatory elements

Gene regulatory regions (also known as ‘ cis-regulatory modules’) in vertebrates are poorly understood and annotated by comparison with protein-coding sequences. The short and degenerate sequences of regulatory elements and their distribution over large intergenic and intronic regions pose a major challenge to genomics scientists. Comparative genomics can be used to identify putative regulatory regions, and to analyse regulatory regions into their constituent transcription factor binding sites. There is need for high throughput assay systems to analyse the function of predicted vertebrate gene regulatory regions

Other applications

Comparative genomics has wide applications in the field of molecular medicine and molecular evolution. The most significant application of comparative genomics in molecular medicine is the identification of drug targets of many infectious diseases. For example, comparative analyses of fungal genomes have led to the identification of many putative targets for novel antifungal. This discovery can aid in target based drug design to cure fungal diseases in human. Comparative analysis of genomes of individuals with genetic disease against healthy individuals may reveal clues of eliminating that disease.

Comparative genomics helps in selecting model organisms. A model system is a simple, idealized system that can be accessible and easily manipulated. For example, a comparison of the fruit fly genome with the human genome discovered that about 60 percent of genes are conserved between fly and human. Researchers have found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. Even more surprisingly, when scientists inserted a human gene associated with early-onset Parkinson’s disease into fruit flies, they displayed symptoms similar to those seen in humans with the disorder, raising the possibility that the tiny insects could serve as a new model for testing therapies aimed at Parkinson’s. Thus, comparative genomics may provide gene functional annotation. Gene finding is an important application of comparative genomics. Comparative genomics identify Synteny (genes present in the same order in the genomes) and hence reveal gene clusters.

Comparative genomics also helps in the clustering of regulatory sites, which can help in the recognition of unknown regulatory regions in other genomes. The metabolic pathway regulation can also be recognized by means of comparative genomics of a species. Dmitry and colleagues have identified the regulons of methionine metabolism in gram-positive bacteria using comparative genomics analysis. Similarly Kai Tan and colleagues have identified regulatory networks of H. influenzae by comparing its genome with that of E. coli. The adaptive properties of organisms like evolution of sex, gene silencing can also be correlated to genome sequence by comparative genomics.

CONCLUSION

The most unexpected finding in comparing the mouse and human genomes lies in the similarities between “ junk” DNA, mostly retro-transposons, (transposons copied from mRNA by reverse transcriptase) in the two species. A survey of the location of retrotransposon DNA in both species shows that it has independently ended up in comparable regions of the genome. Thus “ junk” DNA may have more of a function than was previously assumed. High performance computing tools help in comparing huge genomes. Because of its wide applications and feasibility, automation of comparing genomics is possible. Such Comparisons can aid in predicting the function of numerous hypothetical proteins.

REFERENCES

- en. wikipedia. org/wiki/Comparative genomics
- www. ncbi. nlm. nih. gov
- www. springer. com