Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Comparative genomics - A perspective

Selvarajan Sivashankari[1] and Piramanayagam Shanmughavel[2*]

[1]Department of Bioinformatics, Kongunadu Arts & Science College, Coimbatore - 641029;
[2]Computational Biology Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore - 641046;
Piramanayagam Shanmughavel* - Email: shanvel_99@yahoo.com; * Corresponding author

received November 17, 2006; revised January 26, 2007; accepted February 01, 2007; published online March 27, 2007

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.

Keywords: comparative genomics; genome correspondence; gene identification; genome evolution

Background:

As on Jan 25, 2007, 472 genomes are completely sequenced and yet another 498 are in progress. The rapid progress in genome sequencing demands more comparative analysis to gain new insights into evolutionary, biochemical, genetic, metabolic, and physiological pathways. Comparative genomics is the direct comparison of complete genetic material of one organism against that of another to gain a better understanding of how species evolved and to determine the function of genes and non-coding regions in genomes. It includes a comparison of gene number, gene content, and gene location, the length and number of coding regions (called exons) within genes, the amount of non coding DNA in each genome, and conserved regions maintained in both prokaryotic and eukaryotic groups of organisms. Comparative genomics not only can trace out the evolutionary relationship between organisms but also differences and similarities within and between species. The difference between humans and other organisms can be obtained by comparative investigations. For the purpose of documenting the distinctive features of humans, the most informative research involves comparing humans to our closest relatives, the chimpanzees and apes.

Methodology:

Genome correspondence

Genome correspondence **[1]**, 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-to-many 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., **[2]** developed a method called BBH (Best Bidirectional Hits), which identifies gene pairs that are best matches of each other as orthologous. Tatusov et al., **[3]** 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.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

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. **[4]**

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 **[5]** using comparative genomics and *denovo* motif. The regulatory motifs of the Human Promoters were identified by comparison with other mammals. **[6]** Yet another important finding is the gene and regulatory element by comparison of yeast species. **[7]**

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. **[8]** 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 **[9]** 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 **[10]**, 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 **[11]** have identified the regulons of methionine metabolism in gram-positive bacteria using comparative genomics analysis. Similarly Kai Tan **[12]** and colleagues have identified regulatory networks of *H. influenzae* by comparing its genome with that of *E. coli*. The adaptive properties of organisms **[13]** like evolution of sex, gene silencing can also be correlated to genome sequence by comparative genomics.

Conclusion:

The most unexpected finding in comparing **[14]** 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. **[15]** Such Comparisons can aid in predicting the function of numerous hypothetical proteins.

Acknowledgement:

The authors wish to acknowledge DBT-Bioinformatics Infrastructure Facility, Bharathiar University, Coimbatore 641 046 for providing facilities for their work.

References:

[01] M. Kellis, *et al., J Comput Biol.,* 11:319 (2004) [PMID: 15285895]

[02] W. M. Fitch, *et al., Syst Zool.,* 19:99 (1970) [PMID: 5449325]
[03] W. M. Fitch, *et al., Philos Trans R Soc Lond B Biol Sci.,* 349:93 (1995) [PMID: 8748022]
[04] S. Batzoglou, *et al., Genome Res.,* 10:950 (2000) [PMID: 10899144]
[05] L. Mao & W. J. Zheng, *BMC Bioinformatics,* 7:s21 (2006) [PMID: 17217514]
[06] X. Xie, *et al., Nature,* 434:338 (2005) [PMID:15735639]
[07] M. Kellis, *et al., Nature,* 423:241 (2003) [PMID:12748633]
[08] F. C. Odds, *Rev Iberoam Micol*., 22:229 (2005) [PMID: 16499416]
[09] T. M. Preuss, *Journal of Biomedical Discovery and Collaboration,* 1:17 (2006) [PMID: 17134486]
[10] E. V. Nimwegen, *et al., PNAS,* 99:7323 (2002) [PMID: 12032281]
[11] A. Dmitry, *et al., Nucleic Acids Research,* 32:3340 (2004) [PMID: 15215334]
[12] K. Tan, *et al., Genome Res.,* 11:566 (2001) [PMID: 11282972]
[13] E. Fabre, *et al., Mol Biol Evol.,* 22:856 (2005) [PMID : 15616141]
[14] T. M. Preuss, *Journal of Biomedical Discovery and Collaboration,* 1:17 (2006) [PMID: 17134486]
[15] E. J. Alm, *et al., Genome Res.,* 15:1015 (2005) [PMID:15998914]

Edited by P. Kangueane
Citation: Sivashankari & Shanmughavel, Bioinformation 1(9): 376-378 (2007)