Published on **MacDevCenter** (http://www.macdevcenter.com/)
http://www.macdevcenter.com/pub/a/mac/2004/06/29/bioinformatics.html
See this if you're having trouble printing code examples

Bioinformatics and Comparative Genomics

by Robert Jones
06/29/2004

The complete DNA sequence of the Human Genome is a remarkable achievement for molecular biology and represents the work of many people in a number of large sequencing centers. Far from resting on their laurels, those centers have gone on to sequence the genomes of the mouse, rat, pufferfish, zebrafish, chicken, chimpanzee ... you name it they're sequencing it.

Why this drive to sequence every animal in the zoo? Do we really care about the genetics of pufferfish? In isolation, not so much, but comparisons with the other genomes yield tremendous insights into the genes that are essential for life and those that define the species. They reveal the mechanisms of evolution and the hidden mechanisms of gene regulation.

This article will give a brief introduction to comparative genomics and will show you how to start exploring this treasure trove of data.

A Tale of Two Genomes

Geographic maps are a useful analogy for how we study genomes. If you were given a detailed map of London, you could learn a lot about what defines a large cosmopolitan city. You would see a large number of apartments, shops, and restaurants and might reasonably conclude that these are essential for life in the city. But you could not assess the relative importance of unique features like Buckingham Palace or the Brick Lane street market.

Things would be clearer if you were also given a detailed map of Paris. That too has apartments, shops, and restaurants, confirming your earlier hypothesis. It also has street markets, so perhaps those are an important, albeit secondary, aspect of city life. In contrast, Paris has no "active" royal palaces. Why not? One interpretation might be that Buckingham Palace is an important feature that distinguishes London from other cities. Another might be that a royal family has no function whatsoever in a modern society and survives in London merely as an evolutionary remnant.

It's the same with the human genome. Analysis of the sequence by itself has yielded a vast amount of information. But there are many regions to which we are unable to assign any features. Other regions clearly represent genes but we have no idea what roles the encoded proteins perform in the cell. Some of these will perform vital functions whereas others will be the genetic equivalents of the floppy drive, once important functions that human evolution has rendered obsolete.

Related Reading

Comparing the sequence to a second genome can answer many of these

questions. We can compare one with the other, locate conserved sequence segments and assess their significance. The more genomes we have, the more confident we can become of our assignments and the higher the "resolution" at which we can examine the subtleties.

In choosing the next species to sequence, after the human, one might be tempted to pick a close relative like chimpanzee. But we can learn more from a distant relation like the mouse. As far as biology is concerned, mouse and human are not that different. We both have four limbs and two eyes. We pee, poop, fornicate, and enjoy a nice piece of cheese. The genes responsible for these common structures and functions should be highly conserved and we might expect them to stand out against a background of dissimilar sequences. As we shall see in a moment, that is exactly what happens.

Resolution

The DNA sequence is the most complete representation of a genome that we have available, and it is the form of the data that we actually compare and align, using the BLAST software tools, for example. But we interpret those comparisons in several different ways, which you can think of as multiple resolutions.

First is what we might think of as a medium-resolution view. Take all the known genes from one genome and find their matching genes, if they exist, in the other genome. This gives you a broad picture of which genes have been conserved between species and which are unique.

For each conserved gene, the high-resolution view is the detailed sequence alignment. Translated into the protein sequence, this shows which parts of the protein, encoded by the gene, have been conserved and are, by implication, important for structure and function.

Here we need to introduce the way that genes in "higher" organisms are arranged. Instead of being a single block of sequence that is translated into a protein, as is the case with bacteria, our genes and those of mice, fruit fly, etc. are split into blocks called exons. These are arranged in linear order on the genome with spacer sequences, called introns, in between. The introns can be, and usually are, much longer than the exons. When the gene is turned on, the entire region (introns and exons) is copied into messenger RNA and then the introns are spliced out.

This gene structure of introns and exons varies between species for the same gene, and so that adds another level at which we need to compare genomes. We can think of this as a low-resolution view.

The final level zooms out even further. It describes how groups of genes are arranged between the genomes. This is a concept called synteny.

Synteny

Evolution never makes things simple for biologists. We can't just line up the mouse and human genomes starting at one end of a chromosome and expect to find matching regions one after another. On the time scale of evolution, the process of recombination -- the genetic equivalent of cut-and-paste

-- is continually at work rearranging the genome. Large blocks of genes are moved around within, and between, genomes.

In addition, genetic drift --the appearance and selection of random mutations -- is continually trying to introduce minor changes into the sequence. We know natural selection will conserve important genes but we would not expect, a priori, the arrangement of groups of unrelated genes to be conserved between mouse and human. But Nature always likes to surprise us.

This figure shows the correspondence between regions of human chromosomes on the right, and their counterparts in mouse. Perhaps more striking than the "shuffling" of blocks between the species are the large blocks of sequence that are common to both, and in particular the complete conservation, at this level of resolution, in the sex chromosomes X and Y.

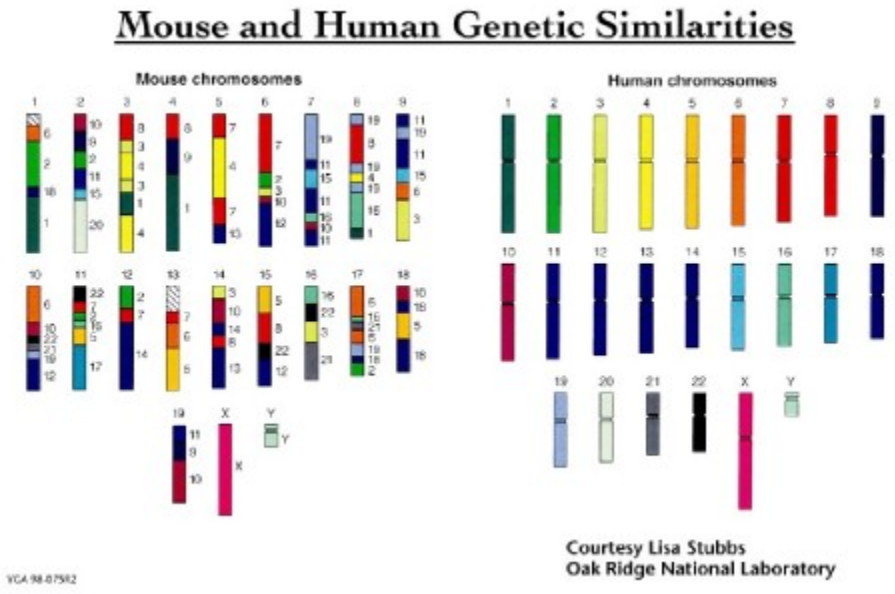


Image credit: U.S. Department of Energy Human Genome Program

Exploring similarities and differences between genomes at all these levels, visualizing them and relating them to other types of information, are just some of the challenges and opportunities facing bioinformatic scientists in this area.

The Software: Genome Browsers

This section introduces you to some of the software tools being used in comparative genomics. Because of the volume and nature of the data involved, almost all the visualization tools in this field use a web interface to access large databases of pre-computed sequence comparisons and annotations.

Although this means the topic doesn't have a real Mac OS X angle to it, I think you will find the following examples interesting. I plan to include hands-on examples throughout this series of articles so you can get a real feel for the data and the software tools we use to explore it. Get in there and get your hands dirty!

To explore comparative genomics we will use the VISTA Genome Browser from Ed Rubin's group at Lawrence Berkeley National Laboratory (LBNL) in Berkeley, Calif. For more information on the

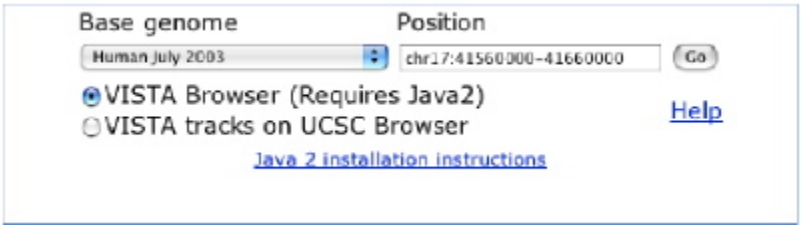
software and its authors, check out the paper listed in the resources section below. It is an excellent way to illustrate some of the ideas discussed here.

The browser is a Java applet that is invoked from your web browser; it requires Java 2. To get started, go to the VISTA Browser home page.

The BRCA1 Gene via VISTA

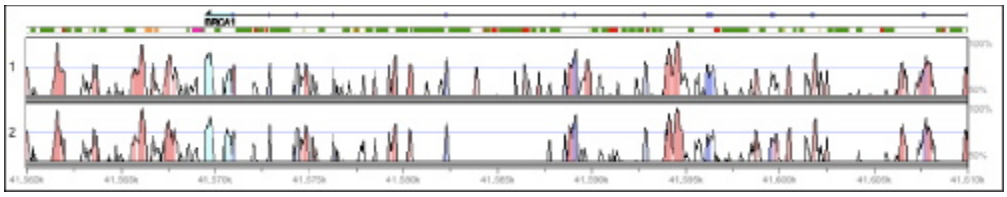
We are going to look at conservation in the BRCA1 gene on human chromosome 17. This is of great medical importance. In the presence of specific mutations, the gene predisposes women to breast cancer. One of the earliest diagnostic tests that came out of the genome project looks for these mutations.

In the Position box on the VISTA home page enter the coordinates of this gene, (chr17:41,560,000-41,660,000) and click Go.



The window that appears shows two "peak and valley" plots of a similar score and in this case the horizontal dimension has been split in the middle to form two panes.

Let's just look at the top pane for now:



There are five tracks in this figure. The x-axis at the bottom shows the sequence coordinates on this chromosome in the human genome. The top track shows the organization of genes that cover this region. In this case we can see from the arrowhead at the left end that BRCA1 is encoded on the reverse strand of the DNA, running right to left.

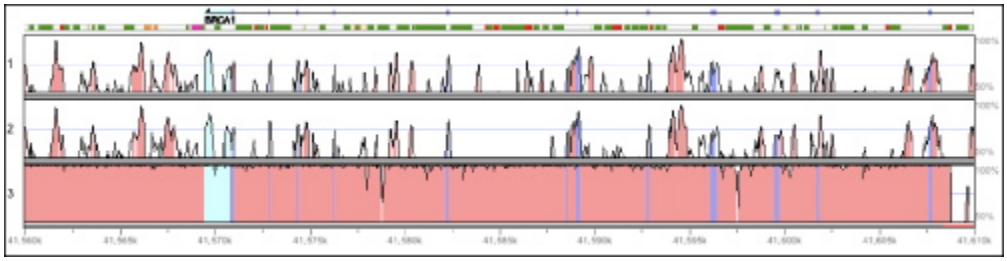
The small blue blocks represent the exons in this gene, separated by the much larger introns. The track below this shows the location of sequence repeats, with the colors representing different types. The genome is packed full of these so-called "junk" sequences, the functions of which are still unclear. Notice that the repeats account for most, but not all, of the intron sequences.

The two large tracks with the peaks and valleys display how well the human sequence is conserved in the genomes of mouse and rat, respectively. The similarity plot is truncated at 50% so any peaks shown indicate strong similarity. These two plots look broadly the same, indicating that the mouse and rat sequences are very similar, as you might expect.

Overall the similarity between these and the human sequence is pretty low but there are many strikingly well-conserved segments, as indicated by the peaks. How these line up with the exons of the human BRCA1 gene is even more striking. All the BRCA1 exons in this region are conserved in mouse and rat. The plots highlight this by shading these segments the same blue color as the exons. This is a strong indication that mouse and rat have a functional equivalent of BRCA1.

But what about the peaks that don't line up with the exons? Those segments with more than 75% similarity are colored pink. Their locations fall in between the human sequence repeats. Evolution has conserved these segments between these distant species for a reason, but what is it? Sequence features like these are the big mystery of the human genome. The presumption is that they are involved in regulation of gene expression or that they play some role in chromosome structure, the higher degrees of packing necessary to fit all this DNA into the cell.

Let's now bring another genome into the comparison and see if that sheds additional light. Go to the Select/Add menu on the left of the applet, select Chimp, and click OK on the dialog that pops up. For clarity I will just show the top panel of the display:



This looks different. With the exception of the region at the right of the panel, and two or three small regions, the chimp sequence is essentially identical to the human. So close, in fact, that a comparison at this resolution is not very informative. This is a good example of why it was important to sequence the genomes from a number of distant species first, in order to identify the conserved regions against a

background of non-conserved sequences. Where the chimpanzee/human comparison has greater value is in the high-resolution comparisons of individual genes, exons, and proteins, as well as in the low-resolution view of synteny.

Here are a few other genes that you might like to look at within the VISTA Browser. You can enter the exact coordinates in the Position box, or just enter the gene name and let VISTA look up the location for you.

HBA1 (chr16:166,679-167,520)
The gene for one of the two protein chains in hemoglobin. HBA1 is small and simple with three exons. It is less than 1,000 nucleotides in length.

MYD88 (ch3:38,140,778-38,145,135)
This gene is involved in the stimulation of cells in the immune system in response to infection. It is relatively small, around 4,000 nucleotides long.

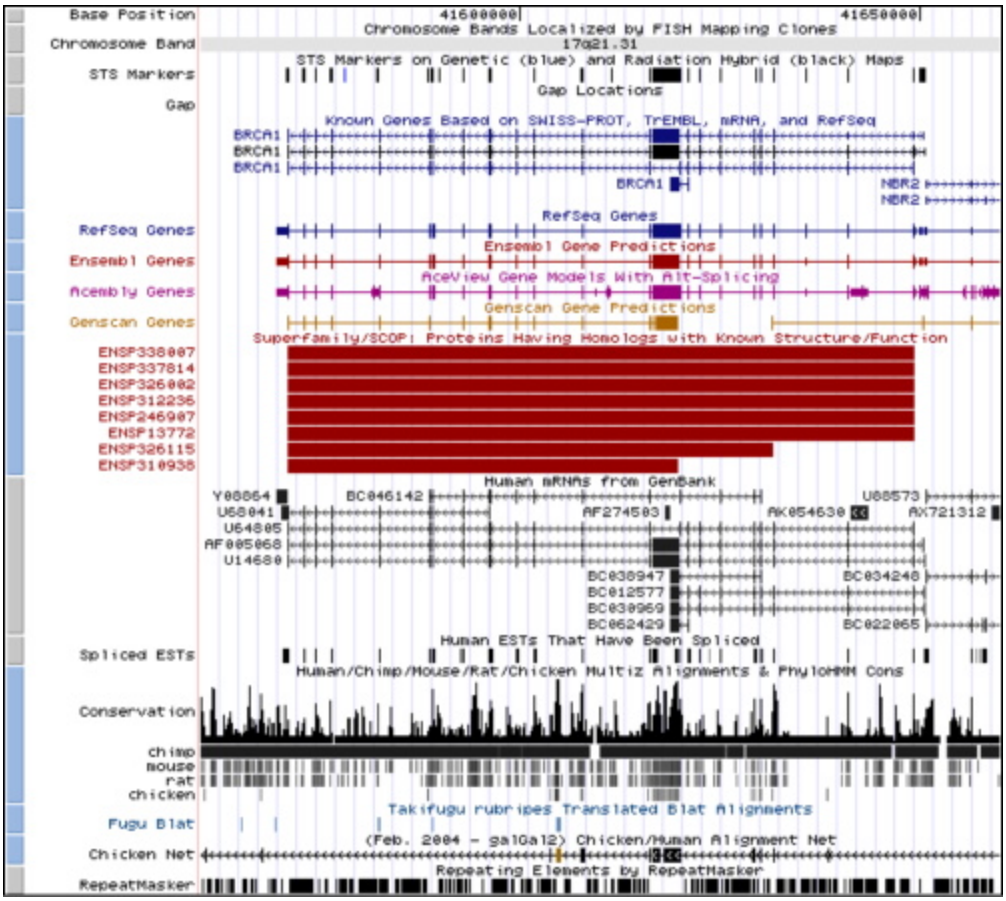
NFKB1 (chr4:103,880,889-103,996,877)
This is an important gene involved in inflammation. The protein encoded by the gene regulates the expression of certain other genes. The exons of NFKB1 are spread over a large segment of the genome, and the gene is about 115,000 nucleotides in length.

Try those genes or just pick a random segment of a chromosome. Play with the interface, zoom in and out, move left and right, and try adding additional genome tracks.

Other Genome Browsers

There are more ways than one to visualize gene organization and conservation. Here are a few other browsers you can try out. Most of these include additional tracks or annotations such as alternative models for gene structure, matches to partial transcript sequences, and other features that I don't have room to describe. Just bear in mind that genome data is not as defined and clear-cut as some of the hype would have you believe.

To illustrate this let's look at BRCA1 in the UC Santa Cruz Genome Browser. Go to the UCSC Genome Browser and enter the BRCA1 coordinates we used earlier.



Yowza. I'm not sure what Edward Tufte would have to say about the design of this, but there are a lot of annotations available and you need to see all of them to evaluate any given region. Many of the features on the image are linked to more-detailed information. Click away and see where it takes you.

For a display that favors tables over images take a look at the Ensembl Viewer from the European Bioinformatics Institute in Cambridge, UK.

The National Center for Biotechnology Information (NCBI) at the NIH has a yet another perspective on visualization with the Map Viewer program. This beast of a URL will show you BRCA1.

Or you can go to the Map Viewer home page and work from there.

The Major Players in Comparative Genomics

Not surprisingly, most of the action in this field is taking place in the big genome centers that generate the data:

- The Broad Institute (nee Whitehead Center for Genome Research), Cambridge, Mass.
- The Sanger Institute, Cambridge, UK.
- Washington University Genome Sequencing Center, St. Louis, Mo.
- Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas.
- DOE Joint Genome Institute, various sites, U.S.
- RIKEN Genomic Sciences Center, Japan.

There is plenty of software development at each of these centers, but as often happens in this field, the bioinformatics have their hands full just managing the data. The integrated browsers described here have all been developed away from the primary data sources at UC Santa Cruz, LBNL in Berkeley, the EBI in the UK, and the NCBI in Washington, D.C.

One area that I've not been able to touch on here is that of comparing bacterial genomes. These sequences are only a few million nucleotides and their genomic organization is much simpler. There are a lot of well-characterized bacterial genomes available, and comparison of these gets really interesting. The Institute for Genomic Research (TIGR) in Maryland and the Sanger Institute in the UK are two of the main centers for this work.

Final Thoughts

You can see from the data shown in the examples given above that comparative genomics is full of challenges. From efficient handling of the vast amounts of data to quantifying similarity from different perspectives, from genome visualization to reverse engineering the course of human evolution, there is a huge amount of work to be done even with the data we already have on disk. Every one of these challenges is an opportunity for developers and creative thinkers to be involved in this remarkable area of science.

I hope this article and its worked examples have given you a taste for this side of bioinformatics and inspires you to learn more. Let me know what you think about it.

Resources

- VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length
  Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) Bioinformatics, 16:1046.
- The UCSC Genome Browser Database
  D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler and W. J. Kent. (2003) Nucleic Acids Research, 31:51-54
- A useful collection of links to databases, genome browsers, etc.
- More links, including those to all the large genome sequencing centers.