Heredity (2003) 91, 533–534

www.nature.com/hdy

NEWS AND COMMENTARY

Comparative genomics

Comparative genomics coming of age

Rebecca F Furlong and Ziheng Yang

Heredity (2003) **91,** 533–534, advance online publication, 22 October 2003; doi:10.1038/sj.hdy.6800372

Without the knowledge of structural and functional features, a genomic sequence is simply a jumble of letters. Comparative analysis of genome sequences from multiple species at different evolutionary distances is fast becoming the predominant approach to identifying functional sequences, such as coding regions and regulatory elements. This approach also takes molecular evolutionary studies to an unprecedented level, providing insights into the forces and mechanisms of the evolutionary process of genes and genomes. A recent study by Thomas *et al* (2003) explores the potential of this comparative approach by sequencing and analysing a genomic region in 12 vertebrate species.

The basic assumption of the approach is that evolutionary conservation implies functional significance. While lack of conservation does not necessarily mean lack of function – indeed, a small subset of genes involved in the immune response and in reproduction are known to be fast evolving, driven by positive Darwinian selection – evolution of most genes and functional regions is dominated by purifying selection; weeding out deleterious mutations. Thus, detection of conserved regions as a means of identifying function has proven very effective, and the power of this approach increases as more genomes are sequenced. However, with the exception of a recent study on yeast (Kellis *et al*, 2003) most previous studies using this approach have been limited to only a few species, such as mouse and human (Mouse Genome Sequencing Consortium 2002) or yeast and human. (Gilligan *et al*, 2002).

Thomas *et al* (2003), however, chose a small genomic region and generated over 12 megabases (Mb) of high-quality sequences from 12 vertebrate species, including mammals, birds, and fish. The targeted genomic region is orthologous to a segment of about 1.8 Mb on human chromosome 7, and encodes 10 genes. One of the genes encodes CFTR, which is mutated in cystic fibrosis, and the whole region is referred to as the 'greater CFTR region'.

To compare genomic sequences between species, a multiple sequence alignment is required. Programs such as ClustalW, familiar to researchers who work with individual genes, are not designed to align genomic data of several Mb in length. This study used blastz for pairwise alignment, which is a gapped blast algorithm modified for long sequences and which, compared with other programs, is sensitive in aligning conserved noncoding regions. Multiple alignments used MultiPipMaker, which is again tailored to deal with long genomic sequences.

The gene number and order were found to be conserved across all 12 species. The amount of noncoding sequences – mainly interspersed repeats – is highly variable among species, causing huge differences in size of the region (from only 162 kb in zebra fish to 1.9 Mb in the human). The methods that the authors used to identify conserved regions appear quite successful, and identified 98% of the exons as well as many noncoding conserved sequences. A large proportion of the latter are known or predicted regulatory elements, while those that are not will be good candidates for further functional studies. The results point to the importance of using multiple species with a range of evolutionary distances for this type of analysis. Comparison between distant species, such as the human and the fish, is effective in identifying coding regions, while many conserved noncoding sequences are identified only in comparison between the mammals. Many conserved regions are missed if only the human and mouse sequences are compared, again highlighting the advantage of including more genomes.

Sequences from multiple vertebrate genomes provide opportunities for refining species phylogenies and for characterising the forces and mechanisms that have shaped the evolutionary process of the genomes. Indeed, the authors identified three transposon insertions shared between primates and rodents, and confirmed the close relationship of primates and rodents, as suggested in a recent phylogenetic analysis based on nucleotide substitutions (Murphy *et al*, 2001).

Less successful is the authors' attempt to characterise the mutation/substitution events from the sequence data, mainly because of lack of suitable statistical methodologies. The authors counted different types of mutation/substitution events, including single-nucleotide mismatches, small indels (<100 BP), large indels (>100 BP), and complex genome-rearrangement events (>100 BP and with inversions or multiple indels). It may appear striking that large indels make the biggest contribution to mutational events in most comparisons. However, this reflects the fact that the authors counted an insertion or deletion of 200 BP as equivalent to 200 single-nucleotide substitutions, and is no evidence that insertions and deletions occur more frequently than substitutions. The overall picture may be somewhat distorted as the authors did not appear to have corrected for multiple hits, and the contribution of single-nucleotide substitutions became increasingly underestimated in more distant comparisons. Pairwise comparison is also expected to be less informative than a simultaneous analysis of all species. In theory, such genomic data should contain information about the relative rates of single-nucleotide substitutions, of insertions and deletions of different sizes, and of genome-rearrangement events, such as inversions and translocations. A probabilistic model is necessary for efficient use of information in the data. Most statistical modelling in molecular sequence evolution has dealt with single-nucleotide substitutions, while most algorithms designed to infer genome-rearrangement events are parsimony based, lack a probabilistic model, and are ill-suited for estimating evolutionary parameters. Development of statistical methods and computational algorithms that can extract the maximum amount of information from genome data remains a serious challenge.

Comparative genomics has tremendous power, and this particular study shows the wealth of information that may be gained from genomic comparisons. However, the study covers only a small region of the vertebrate genome. Genome structure is nonhomogeneous, so other regions will yield different information. Future large-scale analysis is certain to tell us more about the evolutionary, structural and functional aspects of our genome.

R Furlong and Z Yang are at the Department of Biology, University College London, London WC1E 6BT, UK

e-mail: z.yang@ucl.ac.uk

Gilligan P, Brenner S, Venkatesh B (2002). *Gene* **294**: 35–44.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003). *Nature* **423**: 241–254.

Mouse Genome Sequencing Consortium (2002). *Nature* **420**: 520–562.

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ *et al* (2001). *Science* **294**: 2348–2351.

Thomas JW, Touchman JW, Blakesley RW, Bouffard CG, Beckstrom-Sternberg SM, Margulies EH *et al* (2003). *Nature* **424**: 788–793.